REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	January 1999	Annual (17 Dec 97 - 16 Dec 98)

4. TITLE AND SUBTITLE

Malaria Genome Sequencing Project

5. FUNDING NUMBERS

DAMD17-98-2-8005

6. AUTHOR(S)

J. Craig Venter, Ph.D.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

The Institute for Genomic Research
Rockville, Maryland 20850

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION / AVAILABILITY STATEMENT

Approved for public release; distribution unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

The objectives of this 5-year Cooperative Agreement between TIGR and the USAMRMC, were to: **Specific Aim 1**, sequence 3.5 Mb of P. falciparum genomic DNA; **Specific Aim 2**, annotate the sequence; **Specific Aim 3**, release the information to the scientific community. Excellent progress was made towards achievement of these goals. The complete sequence of P. falciparum chromosome 2 (1 Mb) was determined, published in *Science*, and released on the TIGR web site (http://www.tigr.org/tdb/mdb/pfdb/pfdb.html). This is the first malaria chromosome to be sequenced by the Malaria Genome Sequencing Consortium. Many techniques were developed that will facilitate sequencing of the AT-rich P. falciparum genome, including: modification of the sequencing chemistry; development of assembly software and gap closure methods for AT-rich DNA; development of new gene finding software, GlimmerM; construction of a chromosome 2 YAC map and P. falciparum PAC libraries; and, initiation of microarray studies to examine expression of hundreds of genes. The success of this project demonstrates that the extreme AT-richness of the DNA will not prevent sequencing of the entire genome. Malaria researchers will be able to apply this information to the study of *Plasmodium* biology and to development of new drugs and vaccines for against malaria.

14. SUBJECT TERMS	15. NUMBER OF PAGES
Plasmodium falciparum, malaria, genome, chromosome, sequencing, microarray, software	100
	16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	Unlimited

AD_____

COOPERATIVE AGREEMENT NUMBER DAMD17-98-2-8005

TITLE: Malaria Genome Sequence Project

PRINCIPAL INVESTIGATOR: J. Craig Venter, Ph.D.

CONTRACTING ORGANIZATION: The Institute for Genomic Research
Rockville, Maryland 20850

REPORT DATE: January 1999

TYPE OF REPORT: Annual

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

ATTACHMENT 2

FOREWORD

Opinions, interpretations, conclusions and recommendations are those of the author and are not necessarily endorsed by the U.S. Army.

_____ Where copyrighted material is quoted, permission has been obtained to use such material.

_____ Where material from documents designated for limited distribution is quoted, permission has been obtained to use the material.

CWF Citations of commercial organizations and trade names in this report do not constitute an official Department of the Army endorsement or approval of the products or services of these organizations.

_____ In conducting research using animals, the investigator(s) adhered to the "Guide for the Care and Use of Laboratory Animals," prepared by the Committee on Care and Use of Laboratory Animals of the Institute of Laboratory Animal Resources, National Research Council (NIH Publication No. 86-23, Revised 1985).

_____ For the protection of human subjects, the investigator(s) adhered to policies of applicable Federal Law 45 CFR 46.

CWF In conducting research utilizing recombinant DNA, the investigator(s) adhered to the NIH Guidelines for Research Involving Recombinant DNA Molecules.

CWF In the conduct of research involving hazardous organisms, the investigator(s) adhered to the CDC-NIH Guide for Biosafety in Microbiological and Biomedical Laboratories.

Table of Contents

Introduction

Malaria is caused by apicomplexan parasites of the genus *Plasmodium*. It is a major public health problem in many tropical areas of the world, and also affects many individuals and military forces that visit these areas. In 1994 the World Health Organization estimated that there were 300-500 million cases and up to 2.7 million deaths caused by malaria each year, and because of increased parasite resistance to chloroquine and other antimalarials the situation is expected to worsen considerably [1]. These dire facts have stimulated efforts to develop an international, coordinated strategy for malaria research and control [2]. Development of new drugs and vaccines against malaria will undoubtedly be an important factor in control of the disease. However, despite recent progress, drug and vaccine development has been a slow and difficult process, hampered by the complex life cycle of the parasite, a limited number of drug and vaccine targets, and our incomplete understanding of parasite biology and host-parasite interactions.

The advent of microbial genomics, i.e. the ability to sequence and study the entire genomes of microbes, should accelerate the process of drug and vaccine development for microbial pathogens. As pointed out by Bloom, the complete genome sequence provides the "sequence of every virulence determinant, every protein antigen, and every drug target" in an organism [3], and establishes an excellent starting point for this process. In 1995, an international consortium including the National Institutes of Health, the Wellcome Trust, the Burroughs Wellcome Fund, and the US Department of Defense was formed (Malaria Genome Sequencing Project) to sequence the genome of the human malaria parasite Plasmodium falciparum, and later, a second, yet to be determined, species of Plasmodium. Another major goal of the consortium was to foster close collaboration between members of the consortium and other agencies such as the World Health Organization, so that the knowledge generated by the Project could be rapidly applied to basic research and antimalarial drug and vaccine development programs worldwide.

This report describes progress in the Malaria Genome Sequencing Project achieved by The Institute for Genomic Research and the Malaria Program Naval Medical Research Center, under Cooperative Research Agreement DAMD17-98-2-8005, over the 12 month period from Dec. '97 to Dec '98. The specific aims of the work covered under this cooperative agreement were to:

1. Determine the sequence of 3.5 megabases of the *P. falciparum* genome (clone 3D7):

a) Construct small-insert shotgun libraries (1-2 kb inserts) of chromosomal DNA isolated from preparative pulsed-field gels.

b) Sequence a sufficiently large number of randomly selected clones from a shotgun library to provide 10-fold coverage of the selected

c) Construct P1 artificial chromosome (PAC) libraries (inserts up to 20 kb) of chromosomal DNA isolated from preparative pulsed-field gels.

d) If necessary, generate additional STS markers for the chromosome by i) mapping unique-sequence contigs derived from assembly of the random sequences to chromosome, ii) mapping end-sequences from chromosome-specific PAC clones to YACs.

e) Use TIGR Assembler to assemble random sequence fragments, and order contigs by comparison to the STS markers on each chromosome.

f) Close any remaining gaps in the chromosome sequence by PCR and primer-walking using *P. falciparum* genomic DNA or the YAC, BAC, or PAC clones from each chromosome as templates.

2. Analyze and annotate the genome sequence:

a) employ a variety of computer techniques to predict gene structures and relate them to known proteins by similarity searches against databases; identify untranslated features such as tRNA genes, rRNA genes, insertion sequences and repetitive elements; determine potential regulatory sequences and ribosome binding sites; use these data to identify metabolic pathways in *P. falciparum.*

3. Establish a publicly-accessible *P. falciparum* genome database and submit sequences to GenBank.

We are pleased to report that despite encountering formidable technical challenges, excellent progress has been made towards achievement of these goals. A major milestone in malaria research was achieved by the TIGR/NMRC group with the publication in *Science* of the first complete sequence of a malarial chromosome (chromosome 2). A *P. falciparum* genome web site was also established at TIGR which contains all of the published sequence data and annotation, as well as preliminary data for other chromosomes currently being sequenced (http://www.tigr.org/tdb/mdb/pfdb/pfdb.html). In the course of sequencing chromosome 2, the TIGR/NMRC team collaborated with other groups in the development of optical restriction mapping technology for rapid mapping of whole *Plasmodium* chromosomes, construction of a chromosomome 2 YAC map, development of a *Plasmodium* gene finding program, GlimmerM, and construction of *P. falciparum* PAC libraries. In addition, we initiated sequencing of 3 other *P. falciparum* chromosomes, and provided chromosomal DNA for sequencing to other laboratories in the consortium. Finally, in accordance with a modification of the Cooperative Agreement, we initiated studies using microarray technology to gather information relating to the function of novel *Plasmodium* genes identified by the sequencing effort.

Results

Complete nucleotide sequence of P. falciparum chromosome 2 (Specific Aims 1,2,3)

Although sequencing of the AT-rich *Plasmodium* genome presented unique challenges, determination of the complete genome sequences of several microbes suggested that technology had matured to the point that sequencing of the *P. falciparum* genome was feasible. Several groups began working towards this goal, and an international consortium was formed to fund and

coordinate the project [4,5]. As part of this effort, we sequenced chromosome 2 of *P. falciparum* clone 3D7 using a chromosome-specific shotgun sequencing strategy. This approach was selected to avoid the computational and gap closure problems that would arise from application of a whole-genome random shotgun strategy to an AT-rich 30 Mb genome with current technology. In addition, large insert genomic libraries suitable for a directed sequencing strategy were not available because large fragments of *P. falciparum* DNA are prone to deletion and rearrangement in *E. coli*. Small insert shotgun libraries were prepared to minimize the probability of rearrangments, and the techniques used were designed to avoid UV damage and melting of AT-rich DNA.

P. falciparum clone 3D7 was chosen for sequencing because it can complete all stages of the life cycle, was used in a genetic cross [6], and had been used in the Wellcome Trust Malaria Genome Mapping Project [7]. Parasites were grown *in vitro* [8], and parasites released from host cells by acetic acid lysis were embedded in agarose [9]. Chromosomes were resolved on preparative pulsed field gels (1.2% SeaPlaque GTG agarose, BioRad DRIII apparatus,180-250 sec switch time, 120 field angle, 3.7 V/cm for 90 hours at 14 °C), and the chromosome 2 bands from 5 gels were excised, adjusted to 0.3 M sodium acetate to prevent melting of the AT-rich DNA, and digested with agarase. Exposure to UV light was minimized to prevent DNA damage. The DNA was sheared by nebulization and a shotgun library was prepared in pUC18 as described [10], except that treatment with *E. coli* DNA polymerase I was performed (0.5 mM dNTPs, 16 °C for 10 minutes) after the second ligation step to close nicks prior to electroporation. The ligation mixtures were stored at -20 °C, and aliquots were electroporated into DH10B cells and spread on ampicillin diffusion plates. The shotgun library contained 1 x 10^5 recombinants and had an average insert size of 1.6 kb.

Initial sequencing was done with dye-primer chemistry used previously to sequence *H. influenzae* and the other microbial genomes. However, when sequencing *P. falciparum* clones we observed an apparent artifact with the dye-primer chemistry that resulted in runs of G nucleotide base calls to be incorrectly made following long runs of AT-rich sequence. The artifact did not occur when FS+ dye-terminator chemistry was used on the same template DNAs. In addition, the dye-terminator chemistry produced significantly longer sequence reads than the dye-primer chemistry. The rest of the random-phase sequencing was subsequently performed using the dye-terminator chemistry. Because the gel-purified chromosome 2 DNA was only ~85% pure due to co-migration of chromosome 2 DNA with sheared DNA from other chromosomes, and to provide excess coverage to compensate for the expected non-randomness of the shotgun library, 23,768 sequences (equivalent to ~10X coverage) were obtained during the random sequencing phase.

Sequences were assembled using a version of TIGR Assembler that had been extensively modified to assemble the AT-rich and repeat-rich *Plasmodium* sequences. Two modifications to TIGR Assembler [11] reduced assembly time without sacrificing accuracy. TIGR Assembler identifies and aligns overlapping fragments in two steps. The initial step in assembly is to locate all n-mer oligonucleotides shared between fragment pairs. The software views all fragment pairs with a high degree of n-mer similarity as potentially overlapping, and in the second step the Smith-Waterman method is used to align the fragments. In the bacterial genome projects the value of n used was typically 10-12 nucleotides. However, using n=10 with AT-rich *Plasmodium* DNA resulted in incorrect identification of thousands of potential fragment overlaps, so that the program spent an inordinate amount of time attempting to align the spurious matches. Increasing

n from 10 to 32 greatly minimized this problem. In addition, to prevent false merges during the alignment step, TIGR Assembler was modified to ignore 32-mers that were over-represented in the data set.

Six hundred and ten contigs were obtained and the largest contig was 50 kb. Neighboring contigs were identified and ordered by the program GROUPER, which searches for plasmid templates with forward and reverse reads in different contigs (clone links), and for contigs with sequence similarity at their termini (grasta links). Contigs within a group are separated by sequence gaps (gaps which can be closed by primer walking on the templates identified as clone links, or by editing of the termini of contigs with grasta links), and contigs on the ends of groups mark the ends of physical gaps (gaps for which no shotgun clone has been identified). Ten groups of 114 contigs were localized on the chromosome by comparison to STS markers [12]. Closure of physical and sequence gaps used approaches described previously, with a few modifications to compensate for the AT-richness of the DNA. To close the 9 physical gaps in the central region of the chromosome, primers were synthesized to the ends of all groups longer than 2.5 kb. PCR reactions using genomic DNA as template were performed with primers from adjacent groups. PCR reactions were performed with the Expand Long Template PCR System (Boerhinger Mannheim 1681 842). PCR reactions contained 100 ng of genomic DNA and 15 pmol primer (BioServe Technologies) in a final volume of 50 µl. Cycling conditions were 94 °C for 2 min, followed by 10 cycles of 94 °C for 1 min, 50 or 55 °C for 1 min, and 60 °C for 2 min, 20 cycles of 94 °C for 1 min, 50 or 55 °C for 1 min, and 60 °C for 2 min plus 20 sec per cycle, and 1 cycle at 60 °C for 10 min. The 60 °C extension temperature was neccessary for amplification of AT-rich *P. falciparum* DNA [13]. All PCR reactions were done in 96-well format on Perkin Elmer GeneAmp PCR Systems 9600 or 9700. PCR products were purified using the QIAquick PCR Purification Kit (Qiagen 28104) and sequenced using dye-terminator chemistry.This process closed 3 physical gaps immediately, but PCR products from 2 gaps contained very AT-rich sequence which could not be sequenced completely, and remained as sequence gaps. Those physical gaps for which PCR products could not be obtained in the first step were reasoned to be too large for PCR, and to contain one or more of the unlocalized groups. We therefore performed combinatorial PCRs with one primer from the end of a localized group and the second primer from the ends of free groups. Two gaps were closed by the combinatoral strategy. Finally, 1 physical gap was closed after editing and reassembly, and another gap was closed by sequencing a "missing mate:" a plasmid was identified which had one read pointing off the end of the contig, but for which the opposite read that should fall in the gap had failed in sequencing. This template was resequenced and the new sequence provided sufficient information to close the gap. Five methods were used to close sequence gaps. For contigs which overlapped but had not been merged during assembly, editing and resequencing were performed to close the gaps. Also, many sequence gaps were caused by artifacts in dye-primer reactions, particularly in extremely AT-rich areas. These artifacts either prevented merging of overlapping contigs, or resulted in short sequences that did not extend to the neighboring contig. Templates from short or low-quality dye-primer reactions in the vicinity of sequence gaps were identified and resequenced with dye-terminator chemistry; the high-quality sequence provided by the dye-terminator reactions was sufficient to close many gaps. For those gaps that remained, primer walking on plasmid templates linking adjacent contigs was used. There were 5 sequence gaps that could not be closed by the above methods because the sequence was too AT-rich for primer synthesis and walking. To close these gaps, the artificial transposon

AT-2 [14] was inserted into one of the templates spanning each sequence gap, multiple subclones of each template were sequenced, and the sequences were assembled to close the gap.

The coverage criteria were that every position required at least double-clone coverage (or sequence from a PCR product amplified from genomic DNA), and either sequence from both strands or with two different sequencing chemistries. These criteria ensured accuracy of the assembly and base calls. The sequence was edited manually using TIGR Editor, and additional sequencing reactions were performed to improve coverage and resolve sequence ambiguities.

Verification of the assembly.

The widely reported instability of *P. falciparum* DNA in *E. coli* prompted concern as to the ability to clone, sequence, and accurately assemble *P. falciparum* genomic DNA. The coverage criteria used, particularly the requirement for double clone coverage at each position, ensured that rearrangement or chimerism of a single clone did not lead to incorrect assembly of contigs. To independently confirm the colinearity of the assembled sequence and genomic DNA, *Nhe*I and *Bam*HI optical restriction maps of chromosome 2 DNA were prepared in collaboraton with Dr. David Schwartz at New York University (see Optical Mapping of *P. falciparum* chromosomes) and compared with restriction maps predicted from the sequence. The relative error of predicted and observed fragment sizes was 4.3% and 5.8% for the *Nhe*I and *Bam*Hi maps, respectively. The correspondence between the two data sets showed that there were no major rearrangements in the assembled sequence. Further proof of colinearity was obtained by the comparison of the *in silicio* data to a scaffold of YAC-end sequences from chromosome 2-specific YACs isolated from a 3D7 YAC library (see Construction of chromosome 2 YAC map).

General features.

Chromosome 2 of *P. falciparum* (clone 3D7) is 947 kb in length, and has an overall base composition of 80.2% A+T, in agreement with previous estimates for the *P. falciparum* genome. The chromosome contains a large central region encoding single-copy and several tandemly repeated genes, subtelomeric regions containing variant antigen genes (*var*), RIF-1 elements (repetitive interspersed family) [15] and other repeats, and typical eukaryotic telomeres (Figure 1). The terminal 23 kb portions of the chromosome are non-coding and exhibit 77% identity in opposite orientations. The left and right telomeres consist of tandem repeats of the sequence TT(TC)AGGG [16] totaling 1141 and 551 nt, respectively. The sub-telomeric regions do not appear to be composed of repeat oligomers until approximately 12-20 kb internal to the chromosome, where a sequence primarily composed of a previously reported [17,18] 21 bp tandem repeat (rep20) with the consensus ACTAANNTAGGTCTTANNNT was found. This tandem repeat is found exclusively in these regions and occurs 134 and 96 times in the left and right portions of the chromosome, respectively. The chromosome 2 sequence was inspected for repeats of similar abundance and density; no other non-coding DNA repeats similar to rep20 were found. One region occuring in the coding portion of gene PFG0915w (RESA-H3, [19]) generated 36 copies of the peptide consensus VEE{IS}V{AVE}{PE}{STN}. A more complex region composed of sequences that translate into 106 tandemly repeated subunits of 4 distinct polypeptide repeats that range in size of 12-21 amino acids was found in gene PFG0095c (PfEMP3 [20]).

A region with centromere functions could not be identified based on sequence similarity to *S. cerevisiae* centromeres or other eukaryotic centromeres. However, there were several regions of up to 12 kb that were devoid of large open reading frames and which might contain the centromere. Alternatively, mitotic and meiotic centromeric functions may be defined by a

specific higher-order DNA structure and chromatin-associated protein complexes as in other [21]. In addition to the apparent lack of DNA sequences identifiable as centromeric, we were unable to identify sequences that may be involved in the initiation of chromosome replication. Although little is known about chromosomal replication in *Plasmodium*, it is expected to involve many origins of replication as in *S. cerevisiae*, which is the only eukaryote in which replication origins have been clearly defined [22].

Analysis of predicted coding sequences.

The non-redundant (NR) protein sequence database at the National Center for Biotechnology Information (NIH, Bethesda) was searched using the gapped BLAST program. If necessary, database searches were reiterated using the PSI-BLAST program, which constructs position-dependent weight matrices on the basis of alignments generated by BLAST and employs them for subsequent search iterations. Potential coding regions were predicted using GlimmerM, a eukaryotic gene-finding program based on [23]. GlimmerM, was trained on a set of 117 *P. falciparum* sequences taken from Genbank. Gene models based on Glimmer predictions, similarity of ORFs to known proteins, and prediction of putative signal peptides and transmembrane domains were constructed with the Annotator program (Lixin Xhou, TIGR). In cases where a putative gene had no database match and multiple Glimmer predictions of gene structure, the highest scoring model was reported. After the first set of models were inspected, they were added to the training set and GlimmerM was re-trained. These models should be regarded as preliminary until confirmed by other methods. Protein structural features were delineated using an hierarchical scheme implemented in the UniPred program of the SEALS package [24]. Signal peptides were predicted using the SignalP program with the parameters optimized for eukaryotic proteins [25], and transmembrane helices were predicted using PHThtm. Coiled coil domains were predicted using a modification of the COILS program (John Kuzio, NCBI). Regions of low complexity predicted to form non-globular structures were identified using the SEG program with the following parameters: window length 45, trigger complexity 3.4, extension complexity 3 [26]. Multiple sequence alignments were constructed using the CLUSTALW program or the Gibbs-sampling option of the MACAW program. Transfer RNAs were identified with the tRNAscan program [27]. Systematic gene names based on a scheme similar to that devised for the *S. cereviseae* genome [28] were assigned using the convention PF (for *P. falciparum*), a letter for the chromosome (A for chromosome 1, B for chromosome 2, and so forth...), a 3-digit code ordering the genes from left-to-right in increments of 5 (to allow for future modifications), and a letter denoting the coding strand (w or c).

Two hundred and nine protein-encoding genes and a gene for $tRNA^{Glu}$ (Fig 1., Table 1) were predicted on chromosome 2, giving a gene density of one gene per 4.5 kb, which is significantly lower than in yeast (one gene per 2 kb) but somewhat higher than in *C. elegans* (one gene per 7 kb). Of the 209 protein-encoding genes, 43% contained at least one intron. This should be considered an estimate as it cannot be ruled out that some introns were missed by the current gene finding method, or that some of the non-globular inserts detected in *Plasmodium* proteins (see below) are actually introns. The majority of the spliced genes consist of only two or three exons but for two genes 8 exons were predicted. Thus in terms of intron content and gene density, the *Plasmodium* genome, assessed by the analysis of the first completed chromosome sequence, appears to be intermediate between the condensed state of the yeast genome and the intron-rich genomes of multicellular eukaryotes.

The proteins encoded in chromosome 2 (Table 2) fall into three categories: i) proteins with at least one distinct globular domain conserved in species other than *Plasmodium*; ii) proteins belonging to *Plasmodium*-specific families with identifiable structural features and in some cases, known functions; iii), completely uncharacterized proteins which were predominantly non-globular and often large (the term "non-globular" refers to proteins or domains of proteins that do not assume compact, folded structures). Homologs outside *Plasmodium* were detected for 87 out of the 209 predicted proteins (this includes not only proteins in the 1st category but also some in the 2nd category that have unique domain organization so far found only in *Plasmodium* but that contain a conserved domain). Most of the remaining proteins were predicted to consist primarily of non-globular domains (Table 1). Thus the low fraction of proteins containing conserved domains in *Plasmodium* chromosome 2 is due to the enrichment of the *Plasmodium* genome with genes encoding predominantly non-globular proteins. The abundance of non-globular domains or proteins in *Plasmodium* was very unusual; the proportion of non-globular proteins in other eukaryotes such as *S. cerevisiae* (Table 1), *Caenorhabditis elegans*, and humans is approximately half that observed in *Plasmodium*. Remarkably, 13 predicted proteins on chromosome 2 contained large regions (greater than 30 amino acids) with predicted non-globular structure inserted directly into globular domains, a phenomenon so far unique to *Plasmodium.* The non-globular insertions did not exhibit the AT-bias typical of introns and were not flanked by consensus splice sites.

To determine whether the non-globular domains were encoded in mRNA, RT-PCR was performed with primers flanking the non-globular domains in 11 genes from chromosome 2, using total blood stage cDNA as template. RT-PCR was also performed to assess expression of 2 genes encoding predominantly non-globular proteins. In all cases examined, the RT-PCR product was the same size as the corresponding PCR products from genomic DNA, and the sequence of the RT-PCR product matched the genomic DNA sequence (Figure 2A). Thus it appears likely that most if not all predicted non-globular domains in chromosome 2 gene products are expressed. One interesting example of insertion of a non-globular domain into a well-defined globular domain is seen in a 5'-3' exonuclease (Figure 2B). Alignment of the *Plasmodium* sequence with 4 bacterial exonucleases revealed insertion of a 176 amino acid sequence which which was situated in a region corresponding to a sharp turn between a strand and helix in the 3-dimensional structure. These observations seem to reveal a striking flexibility of eukaryotic proteins in terms of accommodating inserts that do not impair protein function and accordingly, may be excluded from the protein core folding. Structural analysis of *Plasmodium* proteins containing non-globular inserts may be valuable for understanding the general principles of protein folding. The propagation of non-globular domains in *Plasmodium* suggests that such proteins provide specific selective advantages to the parasite. The preponderance of the predicted non-globular domains in *Plasmodium* demonstrates the remarkable plasticity and adaptability of the eukaryotic genome.

Evolutionarily conserved proteins.

The conserved proteins encoded in chromosome 2 are a structurally, functionally and phylogenetically diverse set (Table 2). The majority of the conserved proteins show the greatest similarity to eukaryotic homologs or belong to specifically eukaryotic protein families. However, 15 proteins were significantly more similar to bacterial than to eukaryotic homologs, and furthermore, in 4 cases, analysis of the chromosome 2 sequence revealed the first eukaryotic representative of a protein family that is conserved in bacteria. These proteins may have been transferred to the nuclear genome from an organellar, probably plastid genome, after the

divergence of the apicomplexa from the other eukaryotic lineages. Several of these proteins contained likely N-terminal organellar import peptides and are predicted to function within the apicoplast or the mitochondrion. Of particular interest in this regard were 3 genes encoding proteins involved in fatty acid metabolism. 3-ketoacyl-ACP synthase III (FabH) catalyzes the condensation of acetyl-CoA and malonyl-ACP in Type II (dissociated) fatty acid synthase systems. Type II synthase systems are restricted to bacteria and the plastids of plants, confirming previous hypotheses that the *Plasmodium* apicoplast contains metabolic pathways distinct from those of the host [29,30]. Two genes, apparently products of a tandem duplication event, encode another enzyme of fatty acid metabolism, acyl-CoA synthetase; another member of this protein family encoded on a different *Plasmodium* chromosome has been described as an "octapeptide repetitive antigen", on the basis of unique sequence features of its predicted non-globular domain [31]. The expansion of the fatty acid biosynthesis genes in *Plasmodium* may suggest an as yet uncharacterized metabolic process occurring in the parasite.

Given that the Apicomplexa represent a deep branching in the eukaryotic tree, the presence of genes coding for distinct eukaryotic proteins with conserved domain organization is important to ascertain their origin early in the evolution of eukaryotes. The majority of these genes code for proteins that participate in replication, repair, transcription, or translation. These specifically eukaryotic genes that are highly conserved in *Plasmodium* include the origin recognition complex subunit 5; the excision repair proteins ERCC1 and RAD2; several proteins involved in chromatin dynamics, such as orthologs of the superfamily II helicase BRAHMA, DRING protein containing the RING finger domain, and SNW1; RNA-binding proteins, such as the ortholog of *Drosophila* DRIBBLE protein containing the KH domain, and a small nuclear RNP protein containing the RRM domain; and 2 paralogous proteins containing the DHHC finger domain. Furthermore, several typical eukaryotic proteins involved in secretion are encoded in chromosome 2, such as SEC61 g subunit, the coated pit coatamer subunit, and syntaxin, suggesting early emergence of the eukaryotic secretory system.

A remarkable feature detected in chromosome 2 is the expansion of genes coding for DnaJ-like domains. Proteins of the DnaJ superfamily act as cofactors for the HSP70-type molecular chaperones and participate in cellular processes, such as protein folding and trafficking, complex assembly, organelle biogenesis, and initiation of translation [32]. Five proteins containing DnaJ domains are present on chromosome 2, which suggests prominent roles of this domain in the *Plasmodium* life cycle. Two of these consist primarily of the DnaJ domain, whereas the remaining three (two tandem genes in the right subtelomeric region and one in the left subtelomeric region) also have a large non-globular domain. Several proteins containing the DnaJ domain and similar to the three large DnaJ domain-containing proteins encoded in chromosome 2, have been detected on other chromosomes , indicating that this is a large gene family in *Plasmodium*. One of its members has been described as the ring-infected erythrocyte surface antigen (RESA); in fact, this protein has been shown to bind to the cytoplasmic side of the erythrocyte membrane, which suggests that the DnaJ domains perform chaperone-like functions in the formation of specific protein complexes at this location [33-37]. Interestingly, the DnaJ domains in some of the *P. falciparum* proteins contain substitutions in the critical His-Pro-Asp signature required for interaction with HSP-70-type proteins, which may be related to a modification of the typical chaperone function. The actual functions of the chaperone-like domains in *Plasmodium* surface antigens remain to be determined.

Chromosome 2 encodes 90 predicted membrane proteins, and some of these are obvious members of distinct transporter families, such as amino acid and sugar transporters, indicating

that *Plasmodium* has functional transport systems for a number of metabolites, but the lack of even a single ABC transporter is unexpected as these are almost invariably detected in yeast or prokaryotic genomic stretches of similar size.

Another interesting feature of chromosome 2 is the presence of genes for 5 predicted protein kinases belonging to distinct families (e.g. MAP kinases), and a GAF-domain-containing protein that may be involved in cNMP-dependent signaling. A rough extrapolation of the number of protein kinases found on chromosome 2 to the complete *Plasmodium* genome suggests a total of about 150, indicating that the expansion of protein kinases in *Plasmodium* and accordingly, their role in signal transduction and regulation, are at least as prominent in *Plasmodium* as in yeast. This prominence of regulators is in striking contrast to the situation in bacterial pathogens, which appear to have shed most of the regulatory systems, and is likely to be related to the complex life cycle of the parasite. Phosphorylation and dephosphorylation indeed play an important role in the development and sexual differentiation of malaria [38].

Unique Plasmodial protein families.

Chromosome 2 encompasses 5 families of proteins that are unique to *Plasmodium* in terms of their distinct domain organization, though three of them contain domains conserved in other species. The genes comprising these *Plasmodium*-specific families mostly concentrate in the subtelomeric regions of the chromosome. The most abundant family includes proteins that were dubbed rifins, after the RIF-1 repetitive element. RIF-1 elements are found on most chromosomes, are highly transcribed in blood stage parasites, and contain a 1 kb open reading frame but no initiation codon [15]. Eighteen RIF-1 elements were found in the subtelomeric regions of chromosome 2; two of these appeared to be pseudogenes. Inspection of the sequence upstream of each RIF revealed potential exons which encoded predicted signal peptides, and RT-PCR with schizont RNA showed that 1 of 6 rifin genes tested was transcribed (data not shown).The rifin genes encode polypeptides of 27-35 kD with an extracellular domain containing conserved Cys residues which might participate in disulfide bonding, a transmembrane segment, and a short basic C-terminus (Figure 3). Based on the sequence conservation in the extracellular domain, the rifins identified on chromosome 2 can be grouped into two subfamilies, with a high level of conservation within each subfamily but only a limited similarity between the subfamilies. Clusters of rifin genes similar to those on chromosome 2 were detectable also in the telomeric regions of chromosomes 3 and 14 (unpublished observations). A phylogenetic tree analysis of the rifin sequences indicates a direct relationship between some of the rifin genes in chromosome 2 and other chromosomes, suggesting that in the course of *Plasmodium* evolution, the rifin genes have propagated as a cluster (data not shown). If the number of rifins found on chromosome 2 is representative of the other chromosomes, there may be over 500 rifins in the *P. falciparum* genome (~7 % of all protein-coding genes), making it the most abundant gene family. The location and distribution of the rifin genes resembles that of the *var* genes, which may represent 2-3% of plasmodial genes and encode large proteins (PfEMP1) located on the surface of infected red cells that are involved in antigenic variation, cytoadherence, and rosetting [39-42]. Most *var* genes are located in subtelomeric regions, and *var* gene diversity is thought to be generated by recombination between alleles, a process which might be facilitated by the subtelomeric repeats [43]. The abundance and co-localization of rifins with *var* genes in subtelomeric regions suggests that rifins may be involved in similar processes, which is compatible with the presence of a highly variable region in the predicted extracellular domain of the rifins (Fig. 3). Work is in progress to determine where within the parasite and host cell the rifins are expressed.

Two *var* genes were identified on chromosome 2, one in each subtelomeric region and distal to the rifins. Both *var* genes had the typical two-exon structure, and encoded two Duffy binding-like (DBL) domains that were most similar to DD2 var-1 DBL domains 1 and 4. The PFB0010w *var* gene was unusual in that the SVL region was much shorter than in other known var genes. In addition to the two full-length *var* genes, 6 other small ORFs were identified in the subtelomeric regions that had similarity to *var* sequences. Four of these appear to be pseudogenes, but two others resemble the *var* exon II cDNAs reported previously [40].

Another family of membrane-associated proteins, called SERAs (SErine Repeat Antigens), is of interest in that they contain a cathepsin protease-like domain. A cluster of three SERA genes each composed of 4 exons and all transcribed in the same direction (from centromere to telomere) was known to be located on chromosome 2 [44], and at least one has been extensively evaluated for use in blood stage vaccines. The chromosome 2 sequence demonstrates that the three known SERA genes on chromosome 2 are part of an 8-gene cluster. All but one of these 8 genes seem to have a similar 4 exon structure, the last gene at the 3' end of the cluster contains only 3 exons according to structure predictions. Alignment of the 8 SERA proteins revealed conservation of the central and C-terminal regions, the N-termini being more divergent. The first SERA gene described is the only one that contains a stretch of repeated serines, making the generic name SERA quite inappropriate. The protease domain in SERA is remarkable in that 5 out of the 8 copies contain serine instead of cysteine in the active nucleophile position, suggesting that they are serine proteases with a structure typical of cysteine proteases [45]. The expansion of this protease gene family suggests an important function, possibly in merozoite release from schizonts or processing of merozoite surface proteins.

Two copies of another family of surface antigens, typified by MSP-4 [46] were found on the chromosome. An interesting feature of these proteins is the presence of an epidermal growth factor (EGF) module in their extracellular domains. Together with MSP-1, a multi-EGF domain protein encoded on chromosome 3, and two *Plasmodium* sexual stage antigens [47-49], these are the only proteins outside the animal kingdom that contain EGF repeats, and it appears likely that the sequence coding for this domain was hijacked by *Plasmodium* from its animal host. The plasmodial EGF domain may be involved in parasite adhesion to animal cells through homophilic interactions.

In addition to the families of *Plasmodium*-specific proteins, chromosome 2 contains genes for many secreted and membrane proteins that have no identifiable homologs in the databases. Interestingly, one of these genes encodes a protein with a modified thrombospondin domain, and was found to be transcribed in blood stage parasites (data not shown). Other *Plasmodium* proteins that contain thrombospondin domains, such as sporozoite surface protein 2/TRAP and circumsporozoite protein, are involved in parasite invasion of host cells [50-52] and it is tempting to speculate that this protein is involved in binding of infected red cells to host cell ligands.

Optical mapping of P. falciparum chromosomes (added to Specific Aim 1)

Instability of AT-rich *P. falciparum* DNA in E. coli was one of several major technical difficulties faced at the beginning of the sequencing project. Frequent rearrangements or deletions of the cloned *P. falciparum* DNA might have prevented the sequencing of large parts of the genome, or produce an inaccurate assembly of the sequence. Even if these problems had not

been encountered during the sequencing and assembly process, validation of the final sequence was required to ensure that the completed sequence was an accurate representation of the genome.

We initially planned to use a set of sequence markers from the ends of chromosome 2-specific Yeast Artificial Chromosomes for validation of the sequence (see below). It had been reported that YACs containing *P. falciparum* DNA were quite stable, but workers in other fields had noted instability in some YACs. If a proportion of the *P. falciparum* YACS were unstable, they could not be used for validation.

Dr. David Schwartz at the New York University has developed a new technology, called optical restriction mapping, that can generate restriction maps of large DNA molecules very rapidly. Optical mapping does not require cloning of the DNA to be mapped, so that rearrangements that might occur during cloning are avoided. This technology provides a method for sequence validation that is completely independent of the techniques used for generation of the sequence. A collaboration was established with Dr. Schwartz's laboratory to produce an optical map of chromosome 2. Briefly, purified chromosome 2 molecules prepared by Dr. Dan Carucci at NMRC were fixed to glass substrates, digested *in situ* with *Nhe*I or *Bam*HI, stained with YOYO-1, and the order and size of the DNA fragments was obtained by direct observation of a single DNA molecules by fluorescence microscopy. Fragment size was determined by automated fluorescence intensity measurements [53]. The optical restriction maps were compared with restriction maps predicted from the completed chromosome 2 sequence. The relative error of predicted and observed fragment sizes was 4.3% and 5.8% for the *Nhe*I and *Bam*Hi maps, respectively. The correspondence between the two data sets showed that there were no major rearrangements in the assembled sequence, and confirmed that the chromosome 2 sequence had been properly assembled. A manuscript describing these findings has been submitted for publication [54] (see Appendix).

The successful application of the optical mapping approach to sequence validation of chromosome 2 led the Malaria Genome Sequencing Consortium to recommend that Dr. Schwartz's laboratory be funded to generate restriction maps of the entire *P. falciparum* genome. A map of the complete genome was recently determined, and all of the sequencing centers are using the maps for sequence validation. In addition, TIGR and NMRC are collaborating with Dr. Schwartz in developing methods to use the optical restriction maps for gap closure as well as for validation.

Construction of a chromosome 2 YAC map (Specific Aim 1d)

As mentioned previously, our original plan for verification of the sequence involved comparison of the chromosome 2 sequence to a scaffold of YAC end sequences from a 3D7 YAC library. YAC clones were isolated by PCR screening of the YAC library [7] using primers derived from known chromosome 2 STSs [12]. YACs were sized by pulse-field gel electrophoresis. YAC terminal sequences were obtained by digestion of YAC DNA with AluI, DraI, MseI, RsaI, SspI ,or SwaI. DNA fragments were ligated to adaptors and submitted to PCR amplification with a combination of vector- and adaptor-specific primers. DNA sequencing was performed for both strands using primers for the adaptor and for the yeast vector. A YAC map was constructed by aligning the terminal sequences to the chromosome 2 assembly. A manuscript describing these results is in preparation, and further details will be provided in a subsequent report.

Development of a P. falciparum gene finding program (added to Specific Aim 2)

As the chromosome 2 sequence was nearing completion, it became apparent that a new gene finding program would be needed to aid in annotation of the sequence. Several gene finding programs designed to predict genes in organisms such as human and the plant Arabidopsis had been tested on the *P. falciparum* chromosome 2 sequence, and had not been able to accurately detect known chromosome 2 genes. Dr. Stephen Salzberg at TIGR had been involved in the development of the prokaryotic gene finder Glimmer, which is used at TIGR for gene prediction for all of TIGR's bacterial genome projects. Dr. Salzberg offered to modify Glimmer for prediction of genes in *P. falciparum*, and with other members of his laboratory produced a program called GlimmerM that was used for annotation of chromosome 2.

Details of the GlimmerM software are beyond the scope of this report, but a detailed description of the programs is provided in a paper submitted for publication [23] (Appendix). Briefly, the GlimmerM program, when provided with a training set of well-characterized *P. falciparum* genes, constructs a statistical model of coding sequences and donor and acceptor splice sites. New, uncharacterized *P. falciparum* sequence is then analyzed by GlimmerM, and a set of putative gene models is produced. These models are then evaluated by expert annotators in conjunction with other evidence such as database matches, the presence of signal peptides and transmembrane domains, etc., to produce the final gene models reported in the annotation. These models are based on the best available evidence but shuld be confirmed as preliminary until confirmed by other methods. The accuracy of GlimmerM will improve as the training set of well-characterized *P. falciparum* sequences is enlarged, and further refinements to the algorithm are made. GlimmerM will be extremely useful for annotation of the *P. falciparum* genome, and has already been distributed to other members of the Malaria Genome Sequencing Consortium.

Construction of PAC libraries (Specific Aim 1c)

One of the major impediments to research into the molecular biology of malaria parasites is the lack of a E. coli – based cloning vector that can accommodate large (> 25 kb) inserts of *P. falciparum* DNA without deletion or rearrangement. In principle, a large insert library would be useful for sequencing of the *P. falciparum* genome, by providing either a source of clones for sequencing, or as sequence markers to simplify linking of contigs into groups.

Recently, a new E. coli cloning system, called Bacterial Artificial Chromosomes (BACs), that can accommodate DNA inserts greater than 100 kb, has been developed. A derivative of BAC vector, called a PAC (P1 artificial chromosome) vector, has also been developed. The BAC and PAC vectors have been shown to stably maintain "difficult" inserts from a variety of species. We collaborated with Dr. Pieter de Jong at the Roswell Park Cancer Institute, to determine whether *P. falciparum* DNA would also be stable in PAC vectors.

P. falciparum DNA was provided to Dr. de Jong's laboratory, and a PAC library with an average insert size of 8 kb was prepared (work in Dr. de Jong's lab was funded by a grant from the Burroughs Wellcome Fund). The library was screened with chromosome 2 STS markers to identify chromosome 2 clones, and the ends of these clones were sequenced. These sequences were compared with the known chromosome 2 sequence, and 62 clones were identified that had 90% sequence identity to chromosome 2 over 100 bp. Inspection of the data suggested that no rearrangements had occurred in these clones, but this conclusion must be confirmed by further

laboratory analyses. Other studies are underway in Dr. de Jong's lab to determine whether the insert sizes predicted from the sequence agree with those measured experimentally, and whether the clones are stably maintained through extended culture periods. If these clones are stable, the PAC library will be useful for generation of end sequences for linking of contig groups.

Sequencing of P. falciparum chromosome 14 (Specific Aim 1)

Sequencing of chromosome 14 is being funded primarily by a grant from the Burroughs Wellcome Fund; funds from this collaborative agreement are being used to accelerate the sequencing, assist in closure and annotation, develop microarrays for chromosome 14, and facilitate rapid utilization of the sequence data by the DoD vaccine and drug development groups.

Library preparation.

Preparation of a high-quality library of sheared DNA fragments is essential for successful shotgun sequencing of microbial genomes. The strategy adopted for sequencing of the *Plasmodium* genome involves purification of chromosomal DNA on pulsed field gels, followed by construction of shotgun libraries and random sequencing. After determination of the appropriate electrophoresis conditions, chromosome 14 DNA was purified by two rounds of preparative pulsed field gel electrophoresis. The first round was performed in 0.8% chromsomal grade agarose (Figure 4), and the second round was in 0.8% LMP agarose. Conditions were: 1x TAE, 500 sec switch time at a field angle of 106 degrees, 3 V/cm for 48 hours at 14 C. The agarose slices from 10 pulsed field gels (20 48-hour runs!) were equilibrated in 0.3 M sodium acetate, melted at 70C, and digested with agarase. The digested agarose was extracted with phenol and the DNA was precipitated with ethanol. After shearing in a nebulizer, a shotgun library was prepared in pUC18 using the v+i method described previously [55]. Two libraries were prepared with insert sizes of 1.2 - 1.6 kb and 1.6 - 2.0 kb; each library contained 1.6×10^7 recombinants.

High-throughput-sequencing.

During the chromosome 2 project we discovered that dye-primer chemistry produced frequent artifacts with very AT-rich templates, but that dichlororhodamine FS+ dye terminator chemistry gave fewer artifacts and longer read lengths. Consequently, all reactions for chromosome 14 were performed with the dichlororhodamine FS+ dye terminator chemistry. Currently, the success rate for chromosome 14 sequencing reactions is 72% and the average read length is 532 nt; the figures for chromosome 2 were 66% and 507 nt, respectively. Furthermore, inspection of the electropherograms indicates that high quality sequences are being obtained, and there are many fewer indications of the artifacts in AT-rich areas that were problematic with chromosome 2. The artifacts in dye primer reads were a major impediment to assembly and gap closure of chromosome 2; the lack of such artifacts in the chromosome 14 data implies that closure of chromosome 14 should be a simpler process.

Early in the random sequencing phase, test assemblies of the chromosome 14 sequences were performed to assess how closely they compared to theoretical predictions for a completely random library. The data indicate that there are no major problems (i.e. excessive vector contamination, pileups due to gross non-randomness, etc.) with the two chromosome 14 libraries being sequenced. However, contrary to expectations, there seems to be about 15% cross-

contamination of the chromosome 14 DNA with DNA from other chromosomes. This is similar to what was found with chromosome 2, and indicates that the second round of gel-purification did not reduce cross-contamination significantly.

Closure of sequence and physical gaps, and validation of the assembly.

The random sequencing phase was completed in December 1998. 74,292 sequence reads, equivalent to 10X coverage, were obtained. An assembly was performed with the TIGR_Assembler software that had been modified to assemble AT- and repeat-rich *P. falciparum* sequences. This produced 1750 contigs and 3397 singletons. The contigs were released to the public on the TIGR web site (http://www.tigr.org/tdb/mdb/pfdb/pfdb.html) in December 1998.The program grouper was then used to group neighboring contigs and identify physical and sequence gaps in the sequence. Grouper produced 458 groups of contigs totaling 5.5 Mb. As with the chromosome 2 project, we know that the chromosome 14 library contained sequences from other chromosomes. Since chromosome 14 is approx. 3.4 Mb in length, and the groups total 5.5 Mb, approx. 2 Mb of the sequence may not be from chromosome 14.

Closure efforts will focus on those groups of contigs proven to be on chromosome 14 by comparison to a set of chromosome 14 sequence markers, and to groups larger than 5 kb. The contigs from the assembly were compared to more than 70 chromosome 14 markers, which identified 135 contigs and 63 contig groups totaling 2.99 Mb, or 88% of the chromosome. With chromosome 2, groups of contigs were ordered on the chromosome by alignment of the contigs with published STS markers. A similar strategy will be used with chromosome 14. In addition, *Bam* HI and *Nhe* I optical restriction maps of chromosome 14 have been prepared in David Schwartz's laboratory at NYU. Optical restriction maps of chromosome 2 were used as an independent verification of the final assembly. For chromosome 14, optical maps will be used during the closure process to assist in ordering of contigs on the chromosome and sizing of the physical gaps. Thus the information provided by grouper, the optical maps, and comparison of contigs to the chromosome 14 STS markers will be used to map the contigs on the chromosome, and the map will serve as the starting point for the gap closure process.

The techniques to be used for gap closure were developed during the sequencing of several microbial genomes at TIGR, and were proven to work, with some modifications, for closure of chromosome 2. Many of these procedures rely on retrieval of data from a relational database that contains all of the shotgun sequence and assembly data. For sequence gaps these techniques include a) performing long gel runs on templates that are near the ends of contigs and point into the gap, b) primer walking on plasmid templates that span the gaps, c) editing of contig ends to remove vector or low quality sequences that prevented contig merging, and d) for very AT-rich gaps that cannot be closed with the other techniques, the artificial transposon AT-2 (3) will be used to insert primer-binding sites into templates spanning the sequence gaps. The AT-2 transposon was used to close six very AT-rich sequence gaps in chromosome 2 that could not be closed with the other techniques, and to solve a long repeat structure within the PfEMP3 gene. Briefly, an *in vitro* transposition reaction was performed to randomly insert the AT-2 transposon into plasmid templates that spanned the sequence gap. Transposon-containing subclones were identified by selection on ampicillin and trimethoprim, and multiple subclones from each sequence gap were sequenced using transposon-specific primers. The sequences were then assembled to close the gaps.

Physical gaps will be closed using several techniques. First, clones at the ends of contigs which have only one good sequence in the database (i.e. forward or reverse), and for which the missing sequence should fall in the gap (the "missing mate") will be identified, and the missing

sequence reaction for these clones will be repeated. The missing mate sequence will often span the physical gap by itself, or identify an unmapped contig which can be used to close the gap. Second, PCR using genomic DNA and primers complementary to the ends of contigs will be used to amplify fragments spanning physical gaps; the fragments will be sequenced and the sequences assembled to close the gaps. As with chromosome 2, this process will proceed in two phases. In the first phase, primers from contigs predicted to be adjacent on the chromosome will be used in PCR reactions with genomic DNA. This will enable rapid closure of the small physical gaps with a minimum number of PCR reactions. In the second phase, primers from the mapped contigs bordering the remaining physical gaps will be used in PCR reactions with primers from the ends of unmapped contigs larger than 2.5 kb. This will identify unmapped contigs that fall within the larger physical gaps between the mapped contigs. PCR conditions for these reactions were optimized during the chromosome 2 project (long-range PCR with 60C annealing temperature).

The sequence will be evaluated with the program check_coverage to ensure that a) all regions of the assembly are covered by at least two shotgun clones, and b) that every base pair in the sequence has been sequenced in both directions with one chemistry, or in one direction with two chemistries. These criteria ensure that the sequence has been assembled correctly and validate individual base calls. The latter criterion is often satisfied by performing 10% of the sequence reactions with dye-primer chemistry. However, given the frequency of sequence artifacts in AT-rich regions observed with the dye-primer chemistry, this may not be appropriate for *P. falciparum*. As we discovered with chromosome 2, inclusion of sequences containing artifacts in an assembly inhibits contig formation and increases the number of sequence gaps in the assembly and the effort required to close them. Consequently, all chromosome 14 sequencing is being done with dye-terminator chemistry, and late in the random phase the coverage status of the assembly will be assessed. Regions with one-direction coverage will be identified, and additional dye-terminator reactions selected from the database will be performed to convert as many as possible to two-direction coverage. Regions with one-direction coverage that remain will then be re-sequenced with dye-primer chemistry. This process will ensure that the coverage criteria are satisfied and minimize potential assembly problems arising from use of dye-primer chemistry. Finally, the sequence will be edited using the program TIGR_Editor, which displays all gel reads and electropherograms for each base in the sequence. Discrepancies will be noted and additional sequencing reactions will be performed to resolve ambiguities. As a last step to confirm colinearity of the assembled sequence and genomic DNA, restriction maps predicted from the sequence will be compared with the chromosome 14 optical restriction maps described above.

(Note: for the sake of clarity the steps in closure process were presented in chronological order. However, to speed closure several steps can proceed simultaneously; for example, coverage can be checked and improved before gap closure is finished. By working in a parallel fashion chromosome 14 can be completed efficiently).

Annotation.

Elucidation of gene structure will be performed with the program GlimmerM, a eukaryotic gene-finding developed at TIGR specifically for the malaria genome project (see section above). Before the annotation of chromosome 14 begins, GlimmerM will be refined to improve accuracy and the training set will be updated with newly-published sequences, so that a more robust gene-finding tool will be available once the sequence is completed. Predicted coding regions will be searched against the sequence and protein databases using our standard methods.

Repetitive elements and other features will also be identified and annotated. Since many genes will have no database matches, defining the boundaries of genes will be challenging. Most of the software necessary for annotation was tested during the chromosome 2 project, and will require only a few minor modifications for use on chromosome 14 The annotation performed under this grant will by necessity be preliminary. Our goal is to provide a starting point for further biological characterization. We will facilitate public access to the sequence by release of preliminary and finished sequence on the TIGR web site (http://www.tigr.org/tdb/mdb/pfdb/pfdb.html). This will include full text- and sequence-based searching of chromosomes 2 and 14, as well as links to other sources of *P. falciparum* sequence data such as the Sanger Center and Stanford University. Preliminary contigs were released on the TIGR web site in December 1998.

Sequencing of chromosomes 10 and 11 (Specific Aim 1)

Chromosomes 10 and 11 are being sequenced with funding provided by the National Institute for Allergy and Infectious Diseases. Sequencing of chromosome 11 began in December 1998 and should be completed in 1999, so that closure can begin late in 1999. Raw sequence reads of chromosome 11 were released on the TIGR web site in January 1999 and will be updated periodically (http://www.tigr.org/tdb/mdb/pfdb/pfdb.html). Construction of chromosome 10 libraries is underway, and sequencing should begin by mid-1999. Progress on these projects will be updated in next year's annual report.

Preparation of chromosome 12 and 13 DNA for sequencing (Specific Aim 1)

Dr. Daniel Carucci of NMRC provided gel-purified chromosomal DNA to Richard Hyman at Stanford University (chromosome 12) and Daniel Lawson at the Sanger Center (chromosome 13). These investigators have used this material for sequencing of chromosomes funded by the BWF and Wellcome Trust.

We have consulted with Richard Hyman and Eula Fung at Stanford University, who are sequencing chromosome 12, regarding the assembly and gap closure techniques we developed during the chromosome 2 work. In addition, during the chromosome 2 project we obtained many sequences from other parts of the genome due to the co-migration of sheared DNA from other chromosomes with chromosome 2. Sequences (4,563) that were not part of the final chromosome 2 assembly were provided to the Stanford group; these will be examined to determine whether they span any gaps in the chromosome 12 sequence.

Microarray studies (added to Specific Aim 1)

Pilot studies conducted at NMRC in conjunction with TIGR have been initiated to develop DNA microarray technology for the functional analysis of the *P. falciparum* genome. DNA microarrays can be used to examine the expression patterns of thousands of genes simultaneously from two or more RNA samples. These RNA samples may be derived from altered growth conditions, different life stages, and others in order to determine the complement

of genes that may be involved in certain regulatory processes. These pilot studies were carried out using the data generated from the chromosome 2 project.

Chromosome 2-specific DNA microarrays were constructed by identifying individual open reading frames (ORFs) from each of the predicted genes. Forward and reverse oligonucleotide primer sequences were designed using the computer program "Primer 3" from at least one of each of the 209 predicted chromosome 2 genes. In total 245 primer pairs were synthesized.

Each ORF was amplified from *P. falciparum* (clone 3D7) genomic DNA by the polymerase chain reaction (PCR) and verified by agarose gel electrophoresis. The PCR products were purified and resuspended in DMSO prior to arraying. The DNA microarrays were prepared using the Molecular Dynamics Arrayer robot to spot less that 1 nanoliter of each amplified product onto the surface of prepared glass slides. Several test arrays were made using commercially available silanated glass slides or internally prepared poly-L-lysine slides. In general, the poly-L-lysine slides gave the best signal to noise ratio.

Total RNA was prepared from cultured *P. falciparum* (clone 3D7) taken at several time points. The cDNA from each RNA species was differentially labelled with either dUTP-Cy3 or dUTP-Cy5 and hybridized to a DNA microarray at 65° C overnight. After several washes the DNA microarrays were scanned using a ScanArray® 3000 dual color confocal laser system. Fluorescence intensity measurements of each spot were made using the computer program ImaGene™. An example of one experiment comparing the gene expression of two stages of *P. falciparum* blood stage development is shown in Figure 5. Studies are now underway to expand the chromosome-specific DNA microarrays to the nearly completed chromosome 3 (Sanger) and additional chromosome projects as the data become available. In addition, a SyBase relational database is being developed to accommodate the vast quantities of DNA microarray data that will be generated from this project. SyBase was chosen as the DNA microarray relational database so as to provide a seamless integration of data generated from the chromosome 2, 10, 11 and 14 projects at TIGR. A Web interface has also been developed to facilitate data entry and data tracking.

Conclusions

The objectives of this 5-year Cooperative Agreement between TIGR and the USAMRMC were to: **Specific Aim 1**, sequence 3.5 Mb of P. falciparum genomic DNA; **Specific Aim 2**, annotate the sequence; **Specific Aim 3**, release the information to the scientific community. Excellent progress was made towards achievement of these goals. The complete sequence of P. falciparum chromosome 2 (1 Mb) was determined, published in *Science*, and released on the TIGR web site (http://www.tigr.org/tdb/mdb/pfdb/pfdb.html). This is the first malaria chromosome to be sequenced by the Malaria Genome Sequencing Consortium. Many techniques were developed that will facilitate sequencing of the AT-rich P. falciparum genome, including: modification of the sequencing chemistry; development of assembly software and gap closure methods for AT-rich DNA; development of new gene finding software, GlimmerM; construction of a chromosome 2 YAC map and P. falciparum PAC libraries; and, initiation of microarray studies to examine expression of hundreds of genes. The success of this project demonstrates that the extreme AT-richness of the DNA will not prevent sequencing of the entire

genome. Malaria researchers will be able to apply this information to the study of *Plasmodium* biology and to development of new drugs and vaccines for against malaria.

References

1. World malaria situation in 1994. *Weekly Epidemiological Record* **72**, 269-276 (1997).

2. Butler, D., Maurice, J. & O'Brien, C. Briefing malaria. *Nature* **386**, 535-540 (1997).

3. Bloom, B. R. A microbial minimalist. *Nature* **378**, 236 (1995).

4. Hoffman, S. L. *et al.* Funding for malaria genome sequencing. *Nature* **387**, 647 (1997).

5. Gardner, M. J. *et al.* The Malaria Genome Sequencing Project. *Protist* **149**, 109-112 (1998).

6. Walliker, D., Quayki, I., Wellems, T. E. & McCutchan, T. F. Genetic analysis of the human malaria parasite *Plasmodium falciparum*. *Science* **236**, 1661-1666 (1987).

7. Foster, J. & Thompson, J. The *Plasmodium falciparum* genome project: a resource for researchers. *Parasitology Today* **11**, 1-4 (1995).

8. Trager, W. & Jensen, W. Cultivation of malaria parasites. *Nature* **273**, 621-622 (1978).

9. Kemp, D. J. *et al.* Size variation in chromosomes from independent cultured isolates of Plasmodium falciparum. *Nature* **315**, 347-350 (1985).

10. Fleischmann, R. D. *et al.* Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* **269**, 496-512 (1995).

11. Sutton, G. S., White, O., Adams, M. D. & Kerlavage, A. R. TIGR Assembler: a new tool for assembling large shotgun sequencing projects. *Genome Science and Technology* **1**, 9-19 (1995).

12. Lanzer, M., de Bruin, D. & Ravetch, J. V. Transcriptional differences in polymorphic and conserved domains of a completed cloned *P. falciparum* chromosome. *Nature* **361**, 654-657 (1993).

13. Su, X. Z., Wu, Y., Sifri, C. D. & Wellems, T. E. Reduced extension temperatures required for PCR amplification of extremely A+T-rich DNA. *Nucleic Acids Res* **24**, 1574-1575 (1996).

14. Devine, S. E. & Boeke, J. D. Efficient integration of artifical transposons into plasmid targets in vitro: a useful tool for DNA mapping, sequencing, and genetic analysis. *Nucleic Acids Research* **22**, 3765-3772 (1994).

15. Weber, J. L. Interspersed repetitive DNA from *Plasmodium falciparum*. *Mol Biochem Parasitol* **29**, 117-124 (1988).

16. Vernick, K. D. & McCutchan, T. F. Sequence and structure of a *Plasmodium falciparum* telomere. *Mol Biochem Parasitol* **28**, 85-94 (1988).

17. Oquendo, P. *et al.* Characterisation of a repetitive DNA sequence from the malaria parasite, *Plasmodium falciparum. Mol Biochem Parasitol* **18**, 89-101 (1986).

18. Patarapotikul, J. & Langsley, G. Chromosome size polymorphism in *Plasmodium falciparum* can involve deletions of the subtelomeric pPFrep20 sequence. *Nucleic Acids Res* **16**, 4331-4340 (1988).

19. Hernandez, R. R., Hinterberg, K. & Scherf, A. Compartmentalization of genes coding for immunodominant antigens to fragile chromosome ends leads to dispersed subtelomeric gene families and rapid gene evolution in *Plasmodium falciparum. Mol Biochem Parasitol* **78**, 137-148 (1996).

20. Pasloske, B. L. *et al.* Cloning and characterization of a *Plasmodium falciparum* gene encoding a novel high-molecular weight host membrane-associated protein, PfEMP3. *Mol Biochem Parasitol* **59**, 59-72 (1993).

21. Hyman, A. A. & Sorger, P. K. Structure and function of kinetochores in budding yeast. *Annu Rev Cell Dev Biol* **11**, 471-495 (1995).

22. Fangman, W. L. & Brewer, B. J. Activation of replication origins within yeast chromosomes. *Annu Rev Cell Biol* **7**, 375-402 (1991).

23. Salzberg, S. L., Pertea, M., Delcher, A. L., Gardner, M. J. & Tettelin, H. Interpolated Markov models for eukaryotic gene finding. *submitted.*

24. Walker, D. R. & Koonin, E. V. SEALS: a system for easy analysis of lots of sequences. *Ismb* **5**, 333-339 (1997).

25. Nielsen, E., Akita, M., Davila, A. J. & Keegstra, K. Stable association of chloroplastic precursors with protein translocation complexes that contain proteins from both envelope membranes and a stromal Hsp100 molecular chaperone. *Embo J* **16**, 935-946 (1997).

26. Wootton, J. C. & Federhen, S. Analysis of compositionally biased regions in sequence databases. *Methods Enzymol* **266**, 554-571 (1996).

27. Lowe, T. M. & Eddy, S. R. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res* **25**, 955-964 (1997).

28. Mewes, H. W. *et al.* Overview of the yeast genome [published erratum appears in Nature 1997 Jun 12;387(6634):737]. *Nature* **387**, 7-65 (1997).

29. Slabas, A. R. & Fawcett, T. The biochemistry and molecular biology of plant lipid biosynthesis. *Plant Mol Biol* **19**, 169-191 (1992).

30. Wilson, R. J. M., Gardner, M. J., Feagin, J. E. & Williamson, D. H. Have malaria parasites three genomes? *Parasitology Today* **7**, 134-136 (1991).

31. Favaloro, J. M. *et al.* cDNA sequence predicting an octapeptide-repeat antigen of *Plasmodium falciparum*. *Mol Biochem Parasitol* **32**, 297-299 (1989).

32. Cyr, D. M., Langer, T. & Douglas, M. G. DnaJ-like proteins: molecular chaperones and specific regulators of Hsp70. *Trends Biochem Sci* **19**, 176-181 (1994).

33. Bork, P., Sander, C., Valencia, A. & Bukau, B. A module of the DnaJ heat shock proteins found in malaria parasites. *Trends Biochem Sci* **17**, 129 (1992).

34. Coppel, R. L. *et al.* Immune sera recognize on erythrocytes *Plasmodium falciparum* antigen composed of repeated amino acid sequences. *Nature* **310**, 789-792 (1984).

35. Quakyi, I. A. *et al.* Movement of a falciparum malaria protein through the erythrocyte cytoplasm to the erythrocyte membrane is associated with lysis of the erythrocyte and release of gametes. *Infect Immun* **57**, 833-839 (1989).

36. Foley, M., Corcoran, L., Tilley, L. & Anders, R. *Plasmodium falciparum*: mapping the membrane-binding domain in the ring- infected erythrocyte surface antigen. *Exp Parasitol* **79**, 340-350 (1994).

37. Watanabe, J. Cloning and characterization of heat shock protein DnaJ homologues from *Plasmodium falciparum* and comparison with ring infected erythrocyte surface antigen. *Mol Biochem Parasitol* **88**, 253-258 (1997).

38. Bracchi, V., Langsley, G., Thelu, J., Eling, W. & Ambroise, T. P. PfKIN, an SNF1 type protein kinase of Plasmodium falciparum predominantly expressed in gametocytes. *Mol Biochem Parasitol* **76**, 299-303 (1996).

39. Baruch, D. I. *et al.* Cloning the *P. falciparum* gene encoding PfEMP1, a malarial variant antigen and adherence receptor on the surface of parasitized human erythrocytes. *Cell* **82**, 77-87 (1995).

40. Su, Z. *et al.* The large diverse gene family *var* encodes proteins involved in cytoadherence and antigenic variation of Plasmodium falciparum-infected erythrocytes. *Cell* **82**, 89-100 (1995).

41. Smith, J. D. *et al.* Switches in expression of *Plasmodium falciparum var* genes correlate with changes in antigenic and cytoadherent phenotypes of infected erythrocytes. *Cell* **82**, 101-110 (1995).

42. Rowe, J. A., Moulds, J. M., Newbold, C. I. & Miller, L. H. *P. falciparum* rosetting mediated by a parasite-variant erythrocyte membrane protein and complement-receptor 1. *Nature* **388**, 292-295 (1997).

43. Rubio, J. P., Thompson, J. K. & Cowman, A. F. The var genes of Plasmodium falciparum are located in the subtelomeric region of most chromosomes. *Embo J* **15**, 4069-4077 (1996).

44. Knapp, B., Hundt, E., Nau, U. & Kupper, H. A. Molecular cloning, genomic structure and localization in a blood stage antigen of Plasmodium falciparum characterized by a serine stretch. *Mol Biochem Parasitol* **32**, 73-83 (1989).

45. Eakin, A. E., Higaki, J., McKerrow, J. H. & Craik, C. S. Cysteine or serine protease? *Nature* **342**, 132 (1989).

46. Marshall, V. M. *et al.* A second merozoite surface protein (MSP-4) of *Plasmodium falciparum* that contains an epidermal growth factor-like domain. *Infect Immun* **65**, 4460-4467 (1997).

47. Blackman, M. J., Ling, I. T., Nicholls, S. C. & Holder, A. A. Proteolytic processing of the *Plasmodium falciparum* merozoite surface protein-1 produces a membrane-bound fragment containing two epidermal growth factor-like domains. *Mol Biochem Parasitol* **49**, 29-33 (1991).

48. Kaslow, D. C. *et al.* A vaccine candidate from the sexual stage of human malaria that contains EGF-like domains. *Nature* **333**, 74-76 (1988).

49. Duffy, P. E., Pimenta, P. & Kaslow, D. C. Pgs28 belongs to a family of epidermal growth factor-like antigens that are targets of malaria transmission-blocking antibodies. *J Exp Med* **177**, 505-510 (1993).

50. Robson, K. J. *et al.* A highly conserved amino-acid sequence in thrombospondin, properdin and in proteins from sporozoites and blood stages of a human malaria parasite. *Nature* **335**, 79-82 (1988).

51. Cerami, C. *et al.* The basolateral domain of the hepatocyte plasma membrane bears receptors for the circumsporozoite protein of *Plasmodium falciparum* sporzoites. *Cell* **70**, 1021-1033 (1992).

52. Rogers, W. O. *et al.* Characterization of *Plasmodium falciparum* sporozoite surface protein 2. *Proceedings of the National Academy of Sciences, U.S.A.* **89**, 9176-9180 (1992).

53. Samad, A., Huff, E. J., Cai, W. & Schwartz, D. C. Optical mapping: a novel, single-molecule approach to genomic analysis. *Genome Research* **5**, 1-4 (1995).

54. Jing, J. *et al.* Optical mapping of *Plasmodium falciparum* chromosome 2. *Genome Research* **in press** (1999).

55. Gardner, M. J. *et al.* Chromosome 2 sequence of the human malaria parasite *Plasmodium falciparum*. *Science* **282**, 1126-1132 (1998).

56. Thompson, J. D., Higgins, D. G. & Gibson, T. J. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* **22**, 4673-4680 (1994).

57. Kim, Y. *et al.* Crystal structure of Thermus aquaticus DNA polymerase. *Nature* **376**, 612-616 (1995).

Table 1. **Summary of features of *P. falciparum* chromosome 2 and comparison to *S. cerevisiae* chromosome 3.**

*ND, not determined.

†Protein structural features were predicted as described (11).

Description	Number	
	P. f. chr 2	*S. c.* chr 3
Chromosome length (kb)	945	315
G + C content (%)	19.7	38.6
Exons	24.3	40.0
Introns	13.3	ND*
Kilobases per gene	4.50	1.73
Number of predicted protein coding regions	209	171
Number of genes with introns (%)	90 (43)	4 (2.2)
tRNA genes	1	10
Class of proteins†		
Total	209	171
Secreted (%)	22 (11)	11 (6)
Integral membrane (%)	90 (43)	42 (24)
Integral membrane with multiple predicted transmembrane domains (%)	27 (13)	21 (12)
Containing coiled-coil domains (%)	111 (53)	32 (19)
Containing other large compositionally biased regions with predicted non-globular structure (%)	155 (75)	71 (41)
Completely non-globular (%)	17 (8)	6 (3.5)
With detectable homologs in other species	87 (42)	145 (85)

Table 2. Identification of genes on *P. falciparum* chromosome 2.

PF#, systematic name assigned according to a method adapted from *S. cerevisiae* (11).
Description: Name, if known, and prominent features of the gene.
Abbreviations are as follows: euk, eukaryotic; nt, nucleotide; OO, organellar origin; prt, protein; TP, transit peptide.

PF#	Description
Amino acid biosynthesis	
PFB0200c	aspartate aminotransferase
Biosynthesis of cofactors, prosthetic groups, and carriers	
PFB0130w	prenyl transferase
PFB0220w	ubiquinone biosynthesis methyltrans.
Fatty acid and phospholipid metabolism	
PFB0385w	acyl-carrier protein
PFB0410c	phospholipase A2-like a/b fold hydrolase
PFB0505c	3-ketoacyl carr. prt synthase III, FabH (OO, TP)
PFB0685c	ATP-dept. acyl-CoA synthetase (TP)
PFB0695c	ATP-dept. acyl-CoA synthetase (TP)
Purines, pyrimidines, nucleosides, and nucleotides	
PFB0295w	adenylosuccinate lyase (OO)
DNA metabolism	
PFB0160w	ERCC1-like excision repair prt
PFB0180w	prt with 5'-3' exonucl. domain (OO, TP)
PFB0205c	prt with 5'-3' exonucl. domain (Kem-1 family)
PFB0265c	RAD2 endonucl.
PFB0440c	chromatinic RING finger prt, DRING ortholog
PFB0720c	ori. recognition cmplx subunit 5 (ATPase)
PFB0730w	BRAHMA ortholog (DNA helicase superfamily II)
PFB0840w	replication factor C, 40 kDa subunit (replication activat.)
PFB0875c	chromatin-binding prt (SKI/SNW family)
PFB0895c	replication factor C, 140 kDa subunit (ATPase)
Energy metabolism	
PFB0795w	ATP synthase alpha chain
PFB0880w	FAD-dependent oxidoreductase (OO)
Transcription	
PFB0140w	metal binding prt (DHHC domain)
PFB0175c	prt of the MAK16 family
PFB0215c	prt with Egl-like 3'-5' exonucl. domain
PFB0245c	RNA polymerase 16kD subunit, RPB4-like
PFB0255w	RRM type RNA binding prt
PFB0290c	Zn-ribbon transcription factor(TFIIS family)
PFB0370c	RNA-binding prt (KH domain)
PFB0445c	eIF-4A-like DEAD family RNA helicase
PFB0620w	YOU2-like small euk. C2C2 zinc finger prt
PFB0715w	DNA-directed RNA polymerase subunit 2
PFB0725c	metal binding prt (DHHC domain)
PFB0855c	rRNA methylase (SpoU family) (OO, TP)
PFB0860c	RNA helicase
PFB0865w	small nuclear ribonucleoprt. (SNRNP family)
PFB0890c	pseudouridine synthet. (RsuA fam.); 1st euk. member (OO)
Translation and post-translational modification	
PFB0165w	tRNA-Glu
PFB0240w	PINT domain prt (proteasomal subunit)
PFB0260w	PSD2-like 26S proteasomal subunit
PFB0325c	SERA antigen/ protease with active Cys
PFB0330c	SERA antigen/ protease with active Cys
PFB0335c	SERA antigen/ protease with active Cys
PFB0340c	SERA antigen/ protease with active Ser
PFB0345c	SERA antigen/ protease with active Ser
PFB0350c	SERA antigen/ protease with active Ser
PFB0355c	SERA antigen/ protease with active Ser
PFB0360c	SERA antigen/ protease with active Ser
PFB0380c	phosphatase (acid phosphatase family)
PFB0390w	ribosome releasing factor (OO, TP)
PFB0455w	ribosomal prt L37A
PFB0515w	glycosyl transferase (novel euk. family)
PFB0525w	asparaginyl-tRNA synthetase (OO, TP)
PFB0545c	ribosomal prt L7/L12 (OO)
PFB0550w	euk. peptide chain release factor
PFB0585w	Leu/Phe-tRNA prt transferase, 1st euk. member (OO)
PFB0645c	ribosomal prt L13 (OO)
PFB0830w	ribosomal prt S26
PFB0885w	ribosomal prt S30
Regulatory functions	
PFB0150c	Ser/Thr prt kinase
PFB0510w	GAF domain prt (cyclic nt signal transduct.)
PFB0520w	novel prt kinase
PFB0605w	Ser/Thr prt kinase
PFB0665w	Ser/Thr prt kinase
PFB0815w	calcium-dept. prt kinase (C-term. EF hand)
Transport	
PFB0210c	monosaccharide transporter
PFB0275w	membrane transporter
PFB0435c	predicted amine transporter
PFB0465c	membrane transporter
Cell surface	
PFB0010w	var gene
PFB0015c	rifin
PFB0020c	var gene fragment
PFB0025c	rifin
PFB0030c	rifin
PFB0035c	rifin
PFB0040c	rifin
PFB0045c	var gene fragment
PFB0050c	rifin pseudogene
PFB0055c	rifin
PFB0060w	rifin
PFB0065w	rifin
PFB0100c	knob-associated His-rich prt
PFB0300c	merozoite surface antigen MSP-2
PFB0305c	merozoite surface antigen MSP-5 (EGF domain)
PFB0310c	merozoite surface antigen MSP-4 (EGF domain)
PFB0400w	PfS230 paralog (predicted secreted prt)
PFB0405w	transmission blocking target antigen PfS230
PFB0570w	predicted secreted prt (thrombospondin domain)
PFB0760w	Mtn3/RAG1IP-like prt
PFB0915w	RESA-H3 antigen
PFB0955w	rifin
PFB0975c	var gene fragment
PFB1000w	rifin pseudogene
PFB1005w	rifin
PFB1010w	rifin
PFB1015w	rifin
PFB1020w	rifin
PFB1025w	var gene fragment
PFB1030w	var gene fragment
PFB1035w	rifin
PFB1040w	rifin
PFB1045w	var gene fragment
PFB1050w	rifin
PFB1055c	var gene
Other cellular processes	
PFB0085c	prt with DnaJ domain (RESA-like)
PFB0090c	prt with DnaJ domain
PFB0450w	prt translocation complex, sec61 gamma chain
PFB0480w	syntaxin
PFB0500c	RAB GTPase
PFB0595w	prt with DnaJ domain, DNJ1/SIS1 family
PFB0635w	T-complex prt 1 (HSP60 fold superfamily)
PFB0640c	WEB-1 ortholog, WD40
PFB0750w	VPS45-like prt (STXBP/UNC-18/SEC1 family)
PFB0805c	clathrin coat assembly prt
PFB0920w	prt with DnaJ domain (RESA-like)
PFB0925w	prt with DnaJ domain (RESA-like)
Unknown function	
PFB0270w	member family of bacterial prts (OO)
PFB0320c	member hesB fam. (poss. redox activity, OO,TP)
PFB0420w	YgbB prt, 1st euk. member (OO, TP)
PFB0425c	prt of the YMR7 family

Figure 1 Legend. Gene map of *P. falciparum* chromosome 2.

Predicted coding regions are shown on each strand. Exons of protein encoding genes are indicated by rectangles with lines linking rectangles representing introns. The single $tRNA^{Glu}$ gene is indicated by a cloverleaf structure. Genes are color-coded according to broad role categories as shown in the key. Gene identification numbers correspond to those in Table 2. The letters CC and NG followed by numerals indicate the number of predicted coiled-coil and non-globular domains in the proteins, respectively.

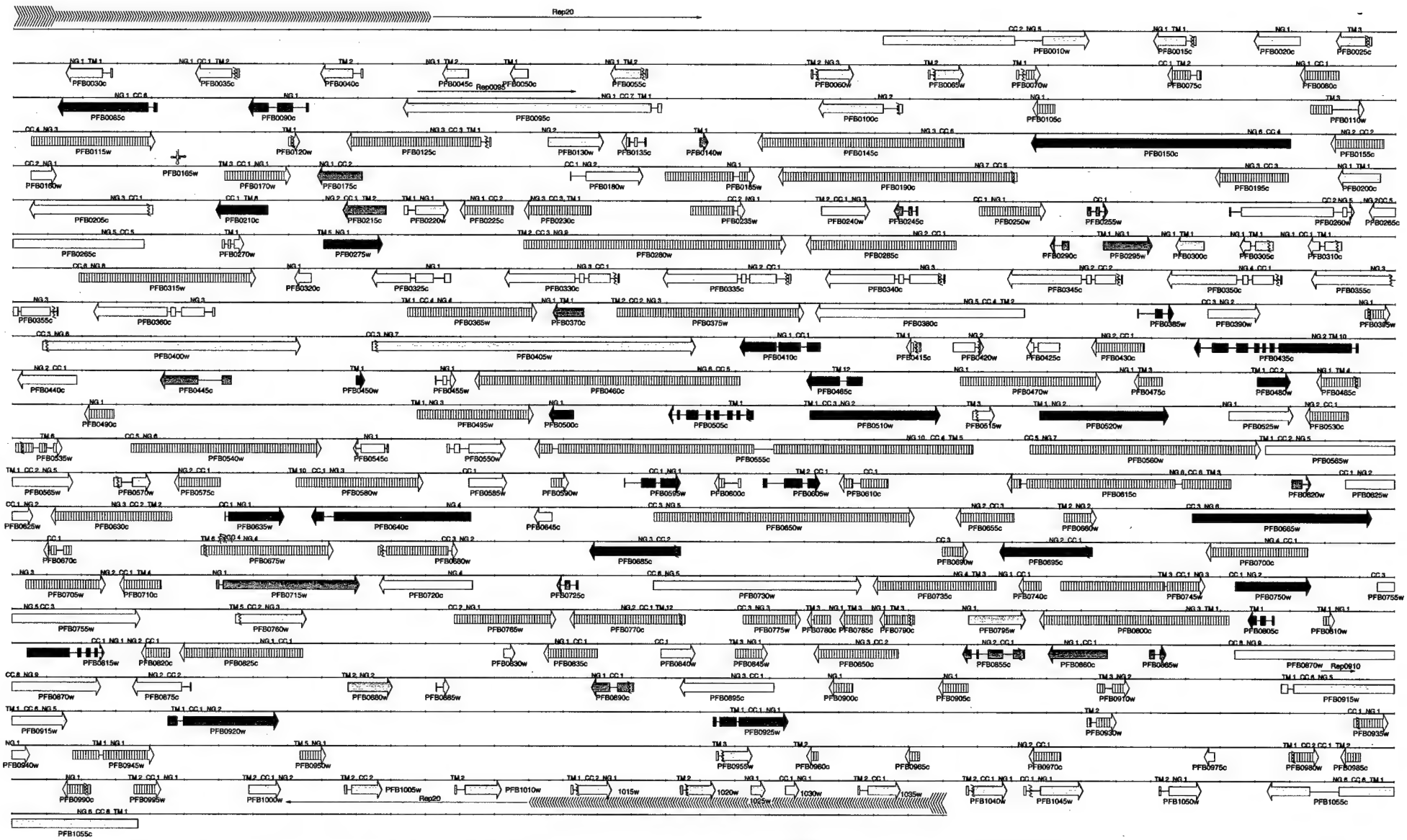

Rep20
Rep0095
Rep0910
Rep20
Amino acid biosynthesis
Biosynthesis of cofactors, prosthetic groups, carriers
Cell surface
Other cellular processes
Energy metabolism
Fatty acid/Phospholipid metabolism
Purines, pyrimidines, nucleosides and nucleotides
Regulatory functions
DNA metabolism
Transport
Translation
Transcription
Hypothetical protein
Unknown function
1 kb
Signal peptide
Coil-Coil domain
Non-globular domain
TM region
Telomeric repeat
Sub-telomeric repeat
Other repeat
tRNA

Figure 1

31

Figure 2 Legend. Confirmation of expression of non-globular domains by RT-PCR.

A. Confirmation of expression of non-globular domains and proteins in blood stages by RT-PCR. Total blood stage RNA (trophozoite or schizont) was amplified by RT-PCR. From left-to-right the samples are PFB0180w (5'-3' exonuclease), PFB0130w (prenyl transferase), PFB0145c (predicted non-globular protein), PFB0265c (RAD2 ortholog), PFB0380c (acid phosphatase), PFB0435c (putative amine transporter), PFB0500c (RAB GTPase), PFB0520w (protein kinase), PFB0525w (asparaginyl-tRNA synthetase), PFB0685c (ATP-dependent acyl CoA synthetase), PFB0720c (origin replication complex subunit 5), PFB0755w (predicted non-globular protein), and PFB0220w (ubiquinone biosynthesis methyltransferase). G, T, and S indicate amplification from 3D7 genomic DNA, trophozoite cDNA, and schizont cDNA, respectively. The PFB0220w gene contains an intron and was used as a control; the primers for this gene spanned the intron, and the smaller size of the products obtained from blood stage RNA confirm that products observed in the RT-PCR reactions were due to amplification from RNA.

B. Multiple alignment of the predicted 5'-3' exonuclease (PFB0180w) encoded in chromosome 2 with homologous bacterial exonuclease domains showing the large non-globular insert in *Plasmodium*. The alignment was constructed using the profile alignment option of CLUSTALW [56]. The alignment column shading is based on a 100% consensus, which is shown underneath the alignment; h indicates hydrophobic residues (A,C,F,I,L,M,V,W,Y; yellow background), u indicates "tiny" residues (G, A, S; green background), o indicates hydroxy residues (S, T), c indicates charged residues (D,E,K,R,H), and "+" indicates positively charged residues (K,R; purple coloring). The aspartates involved in metal coordination are shown by red background and inverse type. Secondary structure elements derived from the crystal structure of *Thermus aquaticus* DNA polymerase [57] are shown above the alignment (H indicates a-helix, and E indicates extended conformation, or b-strand). 5'-3-exo_Aae is a stand alone exonuclease from *Aquifex aeolicus*, and the remaining bacterial sequences are the N-terminal domains of DNA polymerase I.

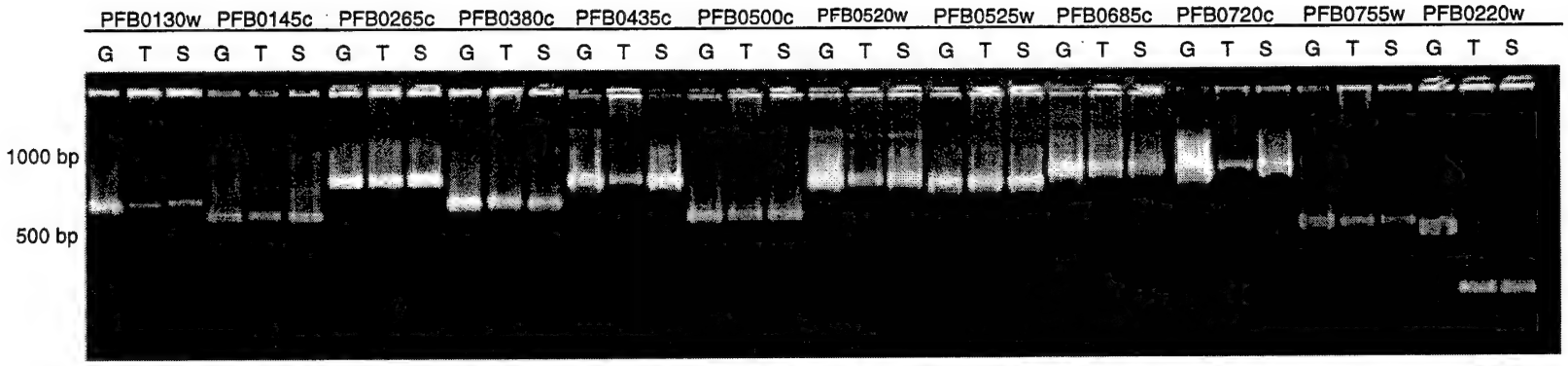

PFB0130w PFB0145c PFB0265c PFB0380c PFB0435c PFB0500c PFB0520w PFB0525w PFB0685c PFB0720c PFB0755w PFB0220w
G T S G T S G T S G T S G T S G T S G T S G T S G T S G T S G T S G T S
1000 bp
500 bp

Figure 2A.

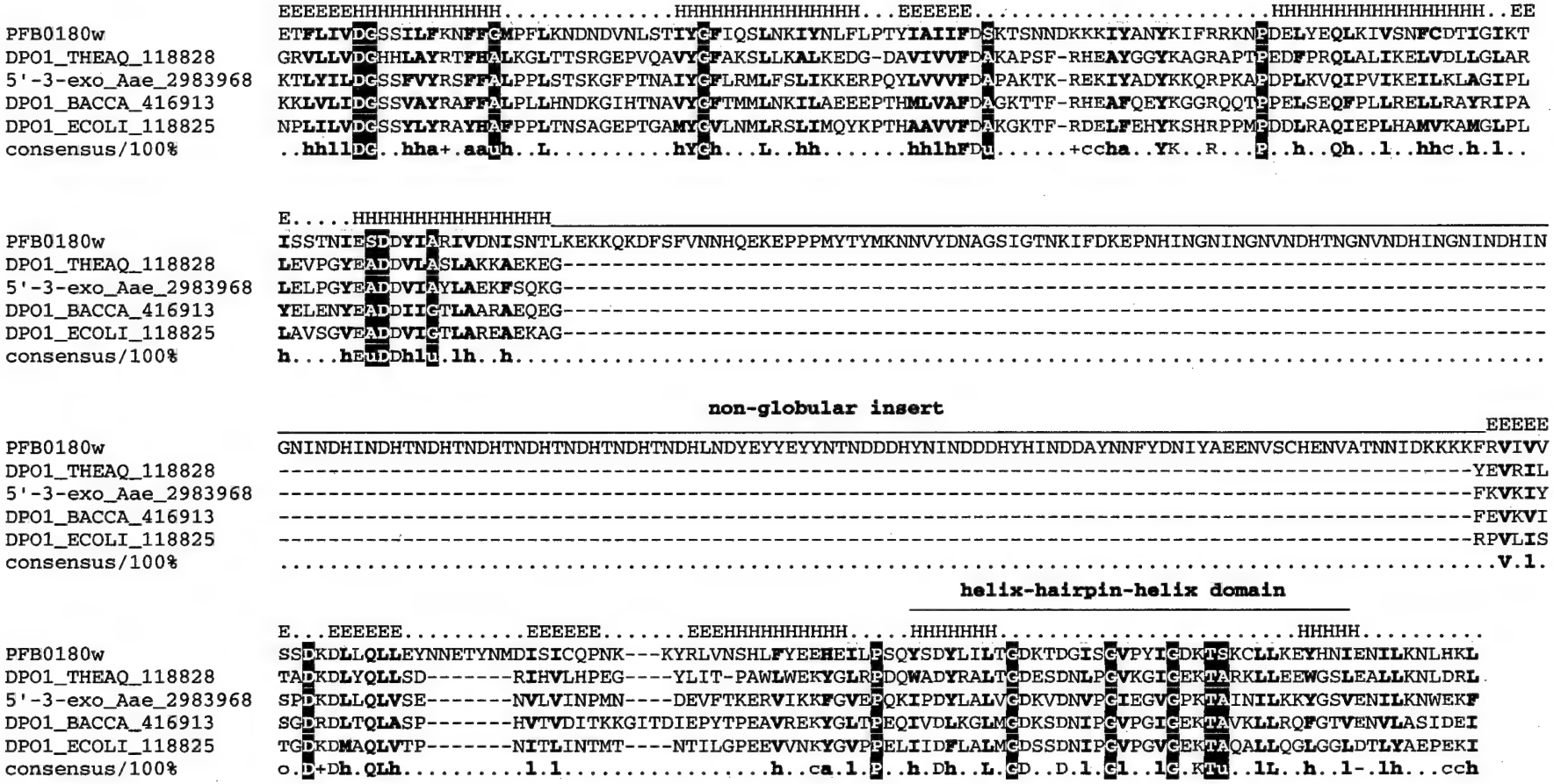

non-globular insert
helix-hairpin-helix domain
PFB0180w
DPO1_THEAQ_118828
5'-3-exo_Aae_2983968
DPO1_BACCA_416913
DPO1_ECOLI_118825
consensus/100%

Figure 2B

Figure 3 Legend. Multiple sequence alignment of rifins encoded on chromosome 2.

The predicted coding regions were aligned with CLUSTALW [56] using the default settings. The alignment column shading is based on a 95% consensus, which is shown underneath the alignment; **h** indicates hydrophobic residues (A,C,F,I,L,M,V,W,Y; yellow background), **p** indicates polar residues (D,E,H,K,N,Q,R,S,T; red coloring), **b** indicates big residues (F,I,L,M,V,W,Y, K,R,Q,E; gray background), and "+" indicates positively charged residues (K,R; red coloring). The cysteines conserved in subsets of rifins are shown by blue shading and inverse coloring.

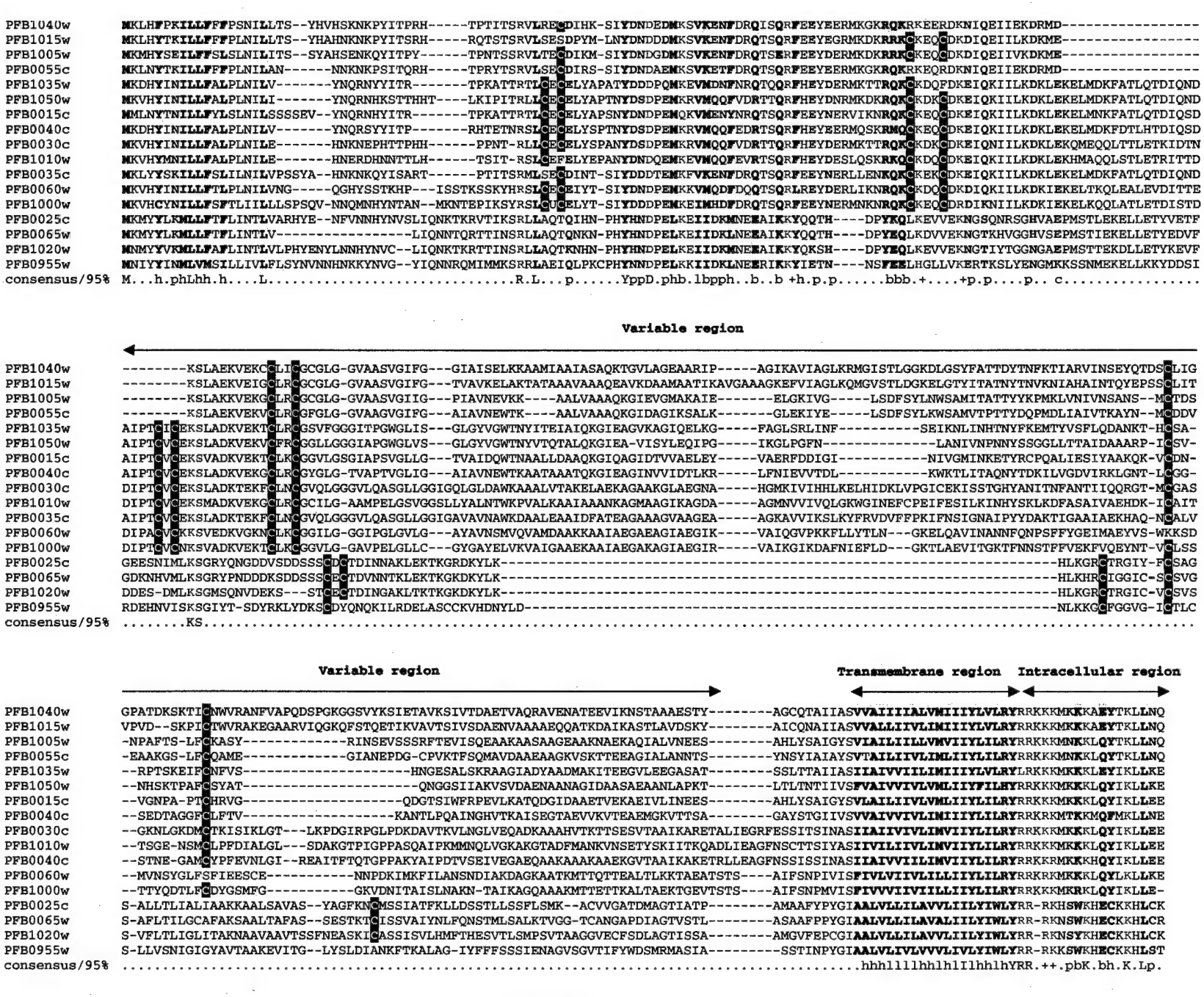

Variable region
Variable region
Transmembrane region
Intracellular region

Figure 3

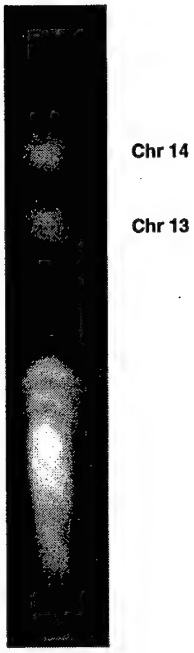


Figure 4. First-round purification of chromosome 14.

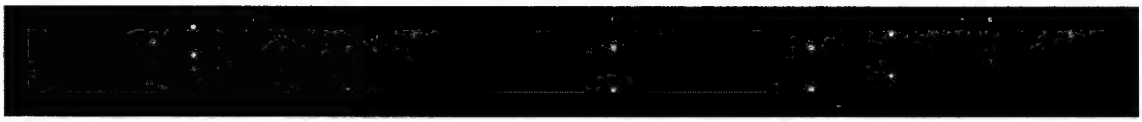


Figure 5. False color image of a chromosome 2 microarray.

Microarray hybridized to *P. falciparum* Cy3-labeled schizont cDNA (green) and Cy5-labeled ring stage cDNA (red).

Appendix

1. Gardner, M. J. *et al.* Chromosome 2 sequence of the human malaria parasite *Plasmodium falciparum*. *Science* **282**, 1126-1132 (1998).

2. Sequencing the malaria protozoan. New York Times, Tuesday, Nov. 6 (1998).

3. Gardner, M. J. *et al.* The Malaria Genome Sequencing Project. *Protist* **149**, 109-112 (1998).

4. Carucci, D. J. *et al.* The malaria genome sequencing project. *Expert Reviews in Molecular Medicine* **http://www-ermm.cbcu.cam.ac.uk/dcn/txt001dcn.htm** (1998).

5. Carucci, D. J. *et al.* Sequencing the genome of *Plasmodium falciparum*. *Current Opinion in Infectious Diseases* **11**, 531-534 (1998).

6. Jing, J. *et al.* Optical mapping of *Plasmodium falciparum* chromosome 2. *Genome Research.* **In press** (1999).

7. Salzberg, S. L., Pertea, M., Delcher, A. L., Gardner, M. J. & Tettelin, H. Interpolated Markov models for eukaryotic gene finding. *Submitted.*

Reprint Series
6 November 1998, Volume 282, pp. 1126–1132

SCIENCE

Chromosome 2 Sequence of the Human Malaria Parasite *Plasmodium falciparum*

Malcolm J. Gardner, Hervé Tettelin, Daniel J. Carucci, Leda M. Cummings, L. Aravind, Eugene V. Koonin, Shamira Shallom, Tanya Mason, Kelly Yu, Claire Fujii, James Pederson, Kun Shen, Junping Jing, Christopher Aston, Zhongwu Lai, David C. Schwartz, Mihaela Pertea, Steven Salzberg, Lixin Zhou,* Granger G. Sutton,† Rebecca Clayton, Owen White, Hamilton O. Smith,† Claire M. Fraser, Mark D. Adams,† J. Craig Venter,† and Stephen L. Hoffman‡

Chromosome 2 Sequence of the Human Malaria Parasite *Plasmodium falciparum*

Malcolm J. Gardner, Hervé Tettelin, Daniel J. Carucci, Leda M. Cummings, L. Aravind, Eugene V. Koonin, Shamira Shallom, Tanya Mason, Kelly Yu, Claire Fujii, James Pederson, Kun Shen, Junping Jing, Christopher Aston, Zhongwu Lai, David C. Schwartz, Mihaela Pertea, Steven Salzberg, Lixin Zhou,* Granger G. Sutton,† Rebecca Clayton, Owen White, Hamilton O. Smith,† Claire M. Fraser, Mark D. Adams,† J. Craig Venter,† Stephen L. Hoffman‡

Chromosome 2 of *Plasmodium falciparum* was sequenced; this sequence contains 947,103 base pairs and encodes 210 predicted genes. In comparison with the *Saccharomyces cerevisiae* genome, chromosome 2 has a lower gene density, introns are more frequent, and proteins are markedly enriched in nonglobular domains. A family of surface proteins, rifins, that may play a role in antigenic variation was identified. The complete sequencing of chromosome 2 has shown that sequencing of the A+T-rich *P. falciparum* genome is technically feasible.

Malaria, a disease caused by protozoan parasites of the genus *Plasmodium*, is one of the most dangerous infectious diseases affecting human populations. Approximately 300 million to 500 million people are infected annually, and 1.5 million to 2.7 million lives are lost to malaria each year, with most deaths occurring among children in sub-Saharan Africa (*1*). Of the four species that cause malaria in humans, *P. falciparum* is the greatest cause of morbidity and mortality. The resistance of the malaria parasite to drugs and the resistance of mosquitoes to insecticides have resulted in a resurgence of malaria in many parts of the world and a pressing need for vaccines and new drugs. The identification of new targets for vaccine and drug development is dependent on the expansion of our understanding of parasite biology; this understanding is hampered by the complexity of the parasite life cycle. The sequencing of the *Plasmodium* genome may circumvent many of these difficulties and rapidly increase our knowledge about these parasites.

The *P. falciparum* genome is ~30 Mb in size; has a base composition of 82% A+T; and contains 14 chromosomes, which range from 0.65 to 3.4 Mb. Chromosomes from different wild isolates exhibit extensive size polymorphism. Mapping studies have indicated that the chromosomes contain central domains that are conserved between isolates and polymorphic subtelomeric domains that contain repeated sequences. *P. falciparum* also contains two organellar genomes. The mitochondrial genome is a 5.9-kb, tandemly repeated DNA molecule; a 35-kb circular DNA molecule, which encodes genes that are usually associated with plastid genomes, is located within the apicoplast [an organelle of uncertain function in *Plasmodium* and the related parasite *Toxoplasma* (*2*)].

Chromosome 2 (GenBank accession number AE001362) was sequenced with the shotgun sequencing approach, which was previously used to sequence several microbial genomes (*3, 4*), with modifications to compensate for the A+T richness of *P. falciparum* DNA (*5*). These modifications included the following: the extraction of DNA from agarose under high-salt conditions to prevent the DNA from melting at a high temperature, the avoidance of ultraviolet (UV) light, the use of the "vector plus insert" protocol for library construction, sequencing with dye-terminator chemistry, the use of a reduced extension temperature in polymerase chain reactions (PCRs), and the use of a transposon-insertion method for the closure of gaps that are very rich in AT. The assembly software was also modified to minimize the misassembly of A+T-rich sequences. The complete sequence included portions of both telomeres and had an average redundancy of 11-fold; colinearity of the final sequence and genomic DNA was proven with optical restriction and yeast artificial chromosome (YAC) maps.

Chromosome 2 of *P. falciparum* (clone 3D7) is 947 kb in length and has an overall base composition of 80.2% A+T. The chromosome contains a large central region that encodes single-copy genes and several duplicated genes, subtelomeric regions that contain variant antigen genes (*var*) (*6–8*), repetitive interspersed family (RIF)–1 elements (*9*) and other repeats, and typical eukaryotic telomeres (Fig. 1). The terminal 23-kb portions of the chromosome are noncoding and exhibit 77% identity in opposite orientations. The left and right telomeres consist of tandem repeats of the sequence TT(TC)AGGG (*10*) and total 1141 and 551 nucleotides (nt), respectively. The subtelomeric regions do not exhibit repeat oligomers until ~12 to 20 kb into the chromosome, where rep20 (*11*) (a 21-bp tandem direct repeat found exclusively in these regions) occurs 134 and 96 times in the left and right ends of the chromosome, respectively. The sequence similarity that was observed between the subtelomeric regions supports previous suggestions that recombination between chromosome ends may be one mechanism by which genetic diversity is generated. A region with centromere functions could not be identified on the basis of sequence similarity to *S. cerevisiae* or other eukaryotic centromeres (*12*). However, several regions of up to 12 kb are devoid of large open reading frames (ORFs) and might contain the centromere. Alternatively, centromeric functions may be defined by higher order DNA structures and chromatin-associated protein complexes (*13*).

Two hundred and nine protein-encoding genes and a gene for $tRNA^{Glu}$ (Fig. 1 and Table 1) were predicted (*14*) on chromosome 2, giving a gene density of one gene per 4.5 kb, which is a value between that observed in yeast (one gene per 2 kb) and in *Caenorhabditis elegans* (one gene per 7 kb). Of the 209 protein-encoding genes, 43% contain at least one intron. This percentage is an estimate

M. J. Gardner, H. Tettelin, L. M. Cummings, S. Shallom, T. Mason, K. Yu, C. Fujii, J. Pederson, K. Shen, L. Zhou, G. G. Sutton, R. Clayton, O. White, H. O. Smith, C. M. Fraser, M. D. Adams, J. C. Venter, The Institute for Genomic Research, 9712 Medical Center Drive, Rockville, MD 20850, USA. D. J. Carucci and S. L. Hoffman, Malaria Program, Naval Medical Research Institute, 12300 Washington Avenue, Rockville, MD 20852, USA. L. Aravind, Department of Biology, Texas A & M University, College Station, TX 70843, USA, and National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA. E. V. Koonin, National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA. J. Jing, C. Aston, Z. Lai, D. C. Schwartz, W. M. Keck Laboratory for Biomolecular Imaging, Department of Chemistry, New York University, New York, NY 10003, USA. M. Pertea, Department of Computer Science, Johns Hopkins University, Baltimore, MD 21218, USA. S. Salzberg, The Institute for Genomic Research, 9712 Medical Center Drive, Rockville, MD 20850, USA, and Department of Computer Science, Johns Hopkins University, Baltimore, MD 21218, USA.

*Present address: ARIAD Pharmaceuticals, 26 Landsdowne Street, Cambridge, MA 02139, USA.
†Present address: Celera Genomics, 45 West Gude Drive, Rockville, MD 20850, USA.
‡To whom correspondence should be addressed. E-mail: hoffmans@nmripo.nmri.nnmc.navy.mil

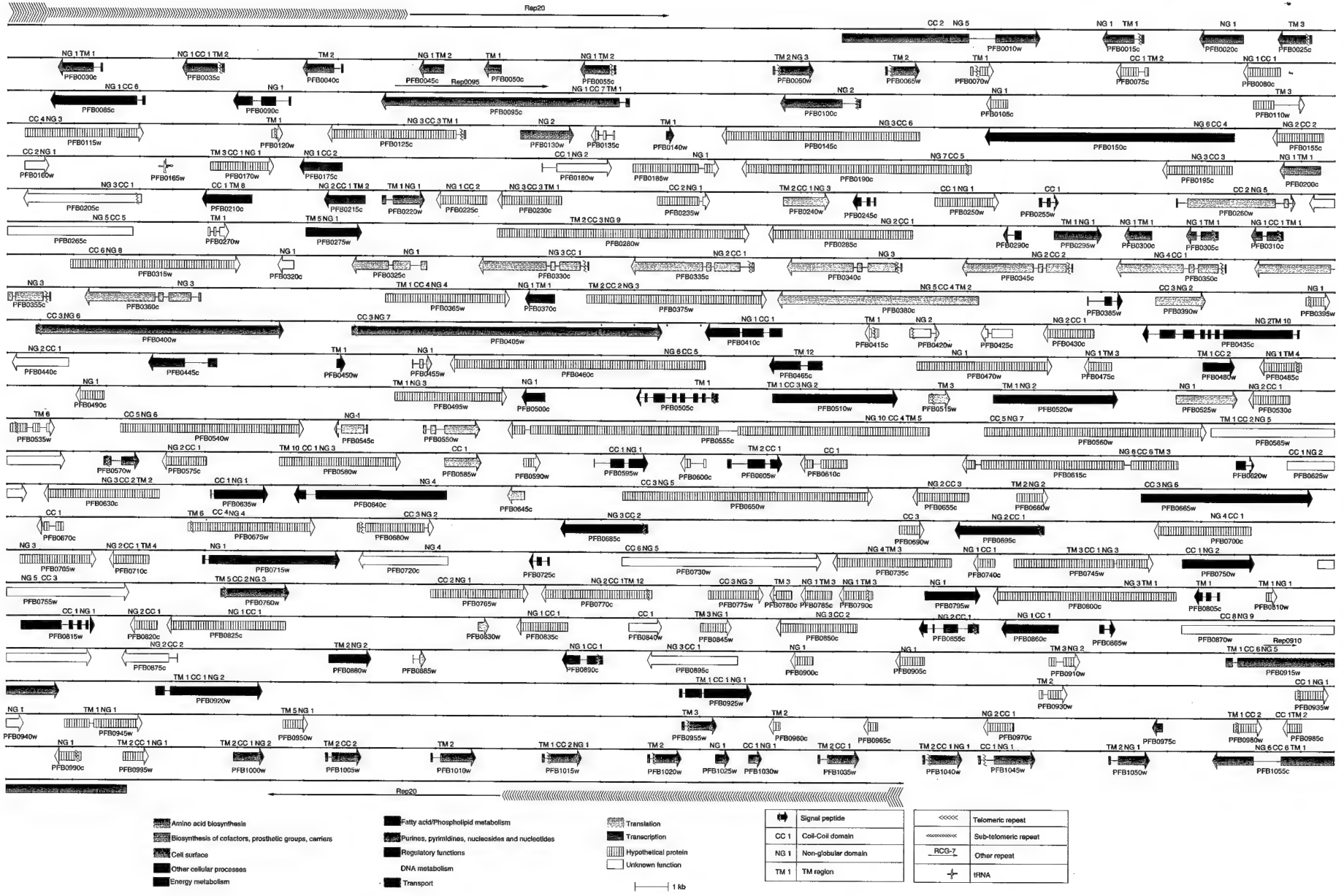


Fig. 1. Gene map of *P. falciparum* chromosome 2. Predicted coding regions are shown on each strand. Exons of protein-encoding genes are indicated by rectangles, and lines linking rectangles represent introns. The single $tRNA^{Glu}$ gene is indicated by a cloverleaf structure. Genes are color-coded according to broad role categories as shown in the key. Genes identification numbers correspond to PF numbers in Table 2. The letters CC, NG, and TM followed by numerals indicate the number of predicted coiled-coil, nonglobular, and transmembrane domains in the proteins, respectively.

because some introns may have been missed by the gene-finding method. Most spliced genes consist of two or three exons. In terms of intron content and gene density, the *Plasmodium* genome, which was assessed by the analysis of the first completed chromosome sequence, appears to be intermediate between the condensed yeast genome and the intron-rich genomes of multicellular eukaryotes.

The proteins encoded in chromosome 2 (Table 2) fall into the following three categories: (i) 72 proteins (34%) are conserved in other genera and contain one or more distinct globular domains; (ii) 47 proteins (23%) belong to *Plasmodium*-specific families with identifiable structural features and, in some cases, known functions; and (iii) 90 predicted proteins (43%) have no detectable homologs, although many contain structural features such as signal peptides and transmembrane domains. Homologs outside *Plasmodium* were detected for 87 (42%) of the 209 predicted proteins. These include proteins in the first category, in addition to those proteins in the second category that possess a conserved domain or domains that are arranged in a manner unique to *Plasmodium*. The percentage of evolutionarily conserved proteins is about two times lower than that found for other genomes, mainly because most of the remaining proteins were predicted to consist primarily of nonglobular domains (*15*) (Table 1). The abundance of nonglobular domains in *Plasmodium* proteins is very unusual; the proportion of proteins with predicted large nonglobular domains in other eukaryotes, such as *S. cerevisiae* (Table 1) or *C. elegans* (*16*), is approximately half that observed in *Plasmodium*. Furthermore, 13 of the 87 conserved proteins on chromosome 2 appear to contain large nonglobular structures (>30 amino acids) that are inserted directly into globular domains, as determined by alignment with homologs from other species.

To determine whether nonglobular domains and proteins are expressed in *P. falciparum*, we performed a reverse transcriptase (RT)–PCR on 11 nonglobular domains and on two genes that encoded predominantly nonglobular proteins, using total blood-stage RNA as a template. In all cases, RT-PCR products were the same size as those that were amplified from genomic DNA, and the sequence of RT-PCR products matched the genomic DNA sequence (*17*). Thus, it is likely that most, if not all, predicted nonglobular domains in chromosome 2 genes are expressed. One example of the insertion of a nonglobular domain into a well-defined globular domain is seen in a protein containing a 5′-3′ exonuclease (Fig. 2). The alignment of the *Plasmodium* sequence with four bacterial exonucleases revealed a 176–amino acid insertion in a region between a strand and a helix in the three-dimensional structure of this protein (*18*). This suggests that eukaryotic proteins can accommodate inserts that may be excluded from the protein core folding without impairing the protein function. The propagation of nonglobular domains in *Plasmodium* suggests that such proteins provide specific selective advantages to the parasite. A structural analysis of *Plasmodium* proteins that contain nonglobular inserts may be valuable for understanding the general principles of protein folding.

Of the 87 conserved proteins that are encoded on chromosome 2, 71 (83%) show the greatest similarity to eukaryotic homologs (Table 2). In contrast, the remaining 16 proteins are most similar to bacterial proteins, and 4 of these represent the first eukaryotic members of protein families that have previously been seen only in bacteria. At least some of these 16 genes may have been transferred to the nuclear genome from an organellar genome after the divergence of the phylum Apicomplexa from other eukaryotic lineages. Several of these proteins appear to contain NH_2-terminal organellar import peptides (*19*) and may function within the apicoplast or the mitochondrion. One such gene encodes 3-ketoacyl–acyl carrier protein (ACP) synthase III (FabH), which catalyzes the condensation of acetyl–coenzyme A and malonyl-ACP in type II (dissociated) fatty acid synthase systems. Type II synthase systems are restricted to bacteria and the plastids of plants, confirming previous hypotheses that the *Plasmodium* apicoplast contains metabolic pathways that are distinct from those of the host (*20*, *21*).

Because the phylum Apicomplexa represents a deep branch in the eukaryotic tree, the presence of eukaryotic-specific genes in *P. falciparum* suggests the appearance of these genes early in eukaryotic evolution. Most of these genes code for proteins that are involved in DNA replication, repair, transcription, or translation (Table 2) and include the origin recognition complex subunit 5, excision repair proteins ERCC1 and RAD2, and proteins involved in chromatin dynamics (such as the BRAHMA helicase, an ortholog of the DRING protein containing the RING finger domain, and chromatin protein SNW1). Furthermore, several eukaryotic proteins involved in secretion are encoded in chromosome 2 (such as the SEC61 γ subunit, the coated pit coatamer subunit, and syntaxin), suggesting an early emergence of the eukaryotic secretory system.

Proteins of the DnaJ superfamily act as cofactors for HSP70-type molecular chaperones and participate in protein folding and trafficking, complex assembly, organelle biogenesis, and initiation of translation (*22*). Five proteins containing DnaJ domains are present on chromosome 2, which suggests multiple roles for this domain in the *Plasmodium* life cycle. Two of these proteins consist primarily of the DnaJ domain, whereas three of the five proteins also contain a large nonglobular domain. Several proteins containing a DnaJ domain have been detected on other chromosomes, indicating that this is a large gene family in *Plasmodium* (*23*). One of its members, the ring-infected erythrocyte surface antigen, binds to the cytoplasmic side of the erythrocyte membrane, suggesting that DnaJ domains perform chaperone-like functions in the formation of protein complexes at this location (*24*). DnaJ domains in some *P.*

Table 1. Summary of features of *P. falciparum* chromosome 2 (*P. f.* chr 2) and comparison to *S. cerevisiae* chromosome 3 (*S. c.* chr 3). Protein structural features were predicted as described (*14*). ND, not determined. Numbers in parentheses indicate the percentage of the total genes or proteins with the specified properties.

Description	Number	
	P. f. chr 2	*S. c.* chr 3
Chromosome length (kb)	947	315
Percent G+C content	19.7	38.6
Exons	24.3	40.0
Introns	13.3	ND
Kilobases per gene	4.50	1.73
Number of predicted protein-coding regions	209	171
Number of genes with introns (%)	90 (43)	4 (2.2)
tRNA genes	1	10
Class of proteins		
Total	209	171
Secreted (%)	22 (11)	11 (6)
Integral membrane (%)	90 (43)	42 (24)
Integral membrane with multiple predicted transmembrane domains (%)	27 (13)	21 (12)
Containing coiled-coil domains (%)	111 (53)	32 (19)
Containing other large compositionally biased regions with predicted nonglobular structure (%)	155 (74)	71 (41)
Completely nonglobular (%)	17 (8)	6 (3.5)
With detectable homologs in other species	87 (42)	145 (85)

Table 2. Identification of genes on *P. falciparum* chromosome 2. The PF number is the systematic name assigned according to a method adapted from *S. cerevisiae* (*14*). The description contains the name (if known) and prominent features of the gene. The table includes genes with homologs in other species and members of *Plasmodium* gene families. An expanded version of this table with additional information is available on the World Wide Web at www.tigr.org/tdb/mdb/pfdb/pfdb.html. Prt, protein; OO, organellar origin; TP, transit peptide; ATP, adenosine triphosphate; euk., eukaryotic; nt, nucleotide.

PF number	Description
Amino acid biosynthesis	
PFB0200c	Aspartate aminotransferase
Biosynthesis of cofactors, prosthetic groups, and carriers	
PFB0130w	Prenyl transferase
PFB0220w	Ubiquinone biosynthesis methyltransferase
Fatty acid and phospholipid metabolism	
PFB0385w	Acyl-carrier prt
PFB0410c	Phospholipase A2-like a/b fold hydrolase
PFB0505c	3-ketoacyl carrier prt synthase III, FabH (OO, TP)
PFB0685c	ATP-dependent acyl-CoA synthetase (TP)
PFB0695c	ATP-dependent acyl-CoA synthetase (TP)
Purines, pyrimidines, nucleosides, and nucleotides	
PFB0295w	Adenylosuccinate lyase (OO)
DNA metabolism	
PFB0160w	ERCC1-like excision repair prt
PFB0180w	Prt with 5′-3′ exonuclease domain (OO, TP)
PFB0205c	Prt with 5′-3′ exonuclease domain (Kem-1 family)
PFB0265c	RAD2 endonuclease
PFB0440c	Chromatinic RING finger prt, DRING ortholog
PFB0720c	Origin recognition complex subunit 5 (ATPase)
PFB0730w	BRAHMA ortholog (DNA helicase superfamily II)
PFB0840w	Replication factor C, 40-kDa subunit (replication activator)
PFB0875c	Chromatin-binding prt (SKI/SNW family)
PFB0895c	Replication factor C, 140-kDa subunit (ATPase)
Energy metabolism	
PFB0795w	ATP synthase alpha chain
PFB0880w	FAD-dependent oxidoreductase (OO)
Transcription	
PFB0140w	Metal-binding prt (DHHC domain)
PFB0175c	Prt of the MAK16 family
PFB0215c	Prt with Egl-like 3′-5′ exonuclease domain
PFB0245c	RNA polymerase 16-kD subunit, RPB4-like
PFB0255w	RRM-type RNA-binding prt
PFB0290c	Zn-ribbon transcription factor (TFIIS family)
PFB0370c	RNA-binding prt (KH domain)
PFB0445c	eIF-4A–like DEAD family RNA helicase
PFB0620w	YOU2-like small euk. C2C2 Zn finger prt
PFB0715w	DNA-directed RNA polymerase subunit 2
PFB0725c	Meta-binding prt (DHHC domain)
PFB0855c	rRNA methylase (SpoU family) (OO, TP)
PFB0860c	RNA helicase
PFB0865w	Small nuclear ribonucleoprt. (SNRNP family)
PFB0890c	Pseudouridine synthetase (RsuA family); first euk. member (OO)
Translation and post-translational modification	
PFB0165w	tRNA-Glu
PFB0240w	PINT domain prt (proteasomal subunit)
PFB0260w	PSD2-like 26S proteasomal subunit
PFB0325c	SERA antigen/protease with active Cys
PFB0330c	SERA antigen/protease with active Cys
PFB0335c	SERA antigen/protease with active Cys
PFB0340c	SERA antigen/protease with active Ser
PFB0345c	SERA antigen/protease with active Ser
PFB0350c	SERA antigen/protease with active Ser
PFB0355c	SERA antigen/protease with active Ser
PFB0360c	SERA antigen/protease with active Ser
PFB0380c	phosphatase (acid phosphatase family)
PFB0390w	Ribosome releasing factor (OO, TP)
PFB0455w	Ribosomal prt L37A
PFB0515w	Glycosyl transferase (novel euk. family)
PFB0525w	Asparaginyl-tRNA synthetase (OO, TP)
PFB0545c	Ribosomal prt L7/L 12 (OO)
PFB0550w	Euk. peptide chain release factor
PFB0585w	Leu/Phe-tRNA prt transferase, first euk. member (OO)
PFB0645c	Ribosomal prt L13 (OO)
PFB0830w	Ribosomal prt S26
PFB0885w	Ribosomal prt S30
Regulatory functions	
PFB0150c	Ser/Thr prt kinase
PFB0510w	GAF domain prt (cyclic nt signal transduction)
PFB0520w	Novel prt kinase
PFB0605w	Ser/Thr prt kinase
PFB0665w	Ser/Thr prt kinase
PFB0815w	Calcium-dependent prt kinase (C-terminus EF hand)
Transport	
PFB0210c	Monosaccharide transporter
PFB0275w	Membrane transporter
PFB0435c	Predicted amino transporter
PFB0465c	Membrane transporter
Cell surface	
PFB0010w	*var* gene
PFB0015c	Rifin
PFB0020c	*var* gene fragment
PFB0025c	Rifin
PFB0030c	Rifin
PFB0035c	Rifin
PFB0040c	Rifin
PFB0045c	*var* gene fragment
PFB0050c	Rifin pseudogene
PFB0055c	Rifin
PFB0060w	Rifin
PFB0065w	Rifin
PFB0100c	Knob-associated His-rich prt
PFB0300c	Merozoite surface antigen MSP-2
PFB0305c	Merozoite surface antigen MSP-5 (EGF domain)
PFB0310c	Merozoite surface antigen MSP-4 (EGF domain)
PFB0400w	PfS230 paralog (predicted secreted prt)
PFB0405w	Transmission-blocking target antigen PfS230
PFB0570w	Predicted secreted prt (thrombospondin domain)
PFB0760w	Mtn3/RAG1IP-like prt
PFB0915w	RESA-H3 antigen
PFB0955w	Rifin
PFB0975c	*var* gene fragment
PFB1000w	Rifin pseudogene
PFB1005w	Rifin
PFB1010w	Rifin
PFB1015w	Rifin
PFB1020w	Rifin
PFB1025w	*var* gene fragment
PFB1030w	*var* gene fragment
PFB1035w	Rifin
PFB1040w	Rifin
PFB1045w	*var* gene fragment
PFB1050w	Rifin
PFB1055c	*var* gene
Other cellular processes	
PFB0085c	Prt with DnaJ domain (RESA-like)
PFB0090c	Prt with DnaJ domain
PFB0450w	Prt translocation complex, SEC61 γ chain
PFB0480w	Syntaxin
PFB0500c	RAB GTPase
PFB0595w	Prt with DnaJ domain, DNJ1/SIS1 family
PFB0635w	T-complex prt 1 (HSP60 fold superfamily)
PFB0640c	WEB-1 ortholog, WD40
PFB0750w	VPS45-like prt (STXBP/UNC-18/SEC1 family)
PFB0805c	Clathrin coat assembly prt
PFB0920w	Prt with DnaJ domain (RESA-like)
PFB0925w	Prt with DnaJ domain (RESA-like)
Unknown function	
PFB0270w	SLR1419 family prt (OO)
PFB0320c	HesB family prt (possible redox activity, OO, TP)
PFB0420w	YgdB prt first euk. member (OO, TP)
PFB0425c	YMR7 family prt

Fig. 2. Multiple alignment of the predicted 5′-3′ exonuclease (PFB0180w) encoded in chromosome 2 with homologous bacterial exonuclease domains showing the large nonglobular insert in *Plasmodium*. The alignment was constructed with the profile alignment option of CLUSTALW (*34*). The alignment column shading is based on a 100% consensus, which is shown underneath the alignment; **h** indicates hydrophobic residues (A, C, F, I, L, M, V, W, and Y), **u** indicates "tiny" residues (G, A, and S), **o** indicates hydroxy residues (S and T), **c** indicates charged residues (D, E, K, R, and H), and **+** indicates positively charged residues (K and R) (*35*). The aspartates involved in metal coordination have a red background and inverse type. Secondary structure elements derived from the crystal structure of *Thermus aquaticus* DNA polymerase (*18*) are shown above the alignment (H indicates α helix, and E indicates extended conformation, or β strand). 5′-3′-exo_Aae is a stand-alone exonuclease from *Aquifex aeolicus*, and the remaining bacterial sequences are the NH_2-terminal domains of DNA polymerase I.

falciparum proteins contain substitutions in the His-Pro-Asp signature that is required for interaction with HSP-70–type proteins, which may indicate a modification of the typical chaperone function.

Chromosome 2 contains five protein families that are unique to *Plasmodium* in terms of their distinct domain organization, although three of them contain domains that are conserved in other genera. The genes encoding the *Plasmodium*-specific families are primarily located near the ends of the chromosome. A single *var* gene was identified in each subtelomeric region. The *var* genes encode large transmembrane proteins (PfEMP1) expressed in knobs on the surface of schizont-infected red cells. PfEMP1 proteins exhibit extensive sequence diversity; are clonally variant; and are involved in antigenic variation, cytoadherence, and rosetting (*6*–*8*). In addition to the full-length *var* genes, six small ORFs were identified in the subtelomeric regions that were similar to *var* sequences. Five of these ORFs resembled the *var* exon II cDNAs or the Pf60.1 sequences that were reported previously (*7*, *25*).

The largest *Plasmodium*-specific family found on chromosome 2 encodes proteins that were dubbed rifins, after the RIF-1 repetitive element. RIF-1 contained a 1-kb ORF but no initiation codon, was found on most chromosomes, and was transcribed in late blood-stage parasites (*9*). The function of the RIF-1 element was unknown. Eighteen ORFs with similarities to RIF-1 were found in the subtelomeric regions of chromosome 2, centromeric to the *var* genes. An inspection of the sequence upstream of these ORFs revealed exons encoding signal peptides, which indicated that the RIF-1 elements were actually genes consisting of two exons. These genes encode potential transmembrane proteins of 27 to 35 kD, with an extracellular domain that contains conserved Cys residues that might participate in disulfide bonding, a transmembrane segment, and a short basic COOH-terminus. The extracellular domain also contains a highly variable region (Fig. 3). RT-PCR with schizont RNA showed that one of six rifin genes that were tested was transcribed. The function of the rifins is unknown, but their sequence diversity, predicted cell surface localization, and expression in schizont stages suggest that, like *var* genes, they may be clonally-variant. Multiple rifin genes were detected in the telomeric regions of chromosomes 3 and 14, suggesting that rifin genes have propagated as clusters in the course of *Plasmodium* evolution (*26*). If the number found on chromosome 2 is representative of other chromosomes, there may be 500 or more rifin genes in the *P. falciparum* genome (~7% of all protein-coding genes), making it the most abundant gene family in this organism. The presence of *var* and rifin genes and other ORFs in subtelomeric regions of *P. falciparum* chromosomes confirms that the subtelomeric regions are not transcriptionally silent (*27*).

Another family of membrane-associated proteins, serine repeat antigens (SERAs), contains a papain protease-like domain. A cluster of three SERA genes, which were all transcribed in the same direction (from centromere to telomere), was known to be on chromosome 2 (*28*); at least one SERA has been evaluated for use in blood-stage vaccines. These genes are part of an eight-gene cluster; seven genes have a similar four-exon structure, but the gene at the 3′ end of the cluster contains only three exons. The protease domains in these proteins are unusual because five of the eight contain serine instead of cysteine in the active nucleophile position, suggesting that they are serine proteases with a structure that is typical of cysteine proteases (*29*).

Two proteins (MSP-4 and MSP-5) that contain an epidermal growth factor (EGF) module in their extracellular domains were identified (*30*, *31*). In organisms that are not classified in the animal kingdom, MSP-4,

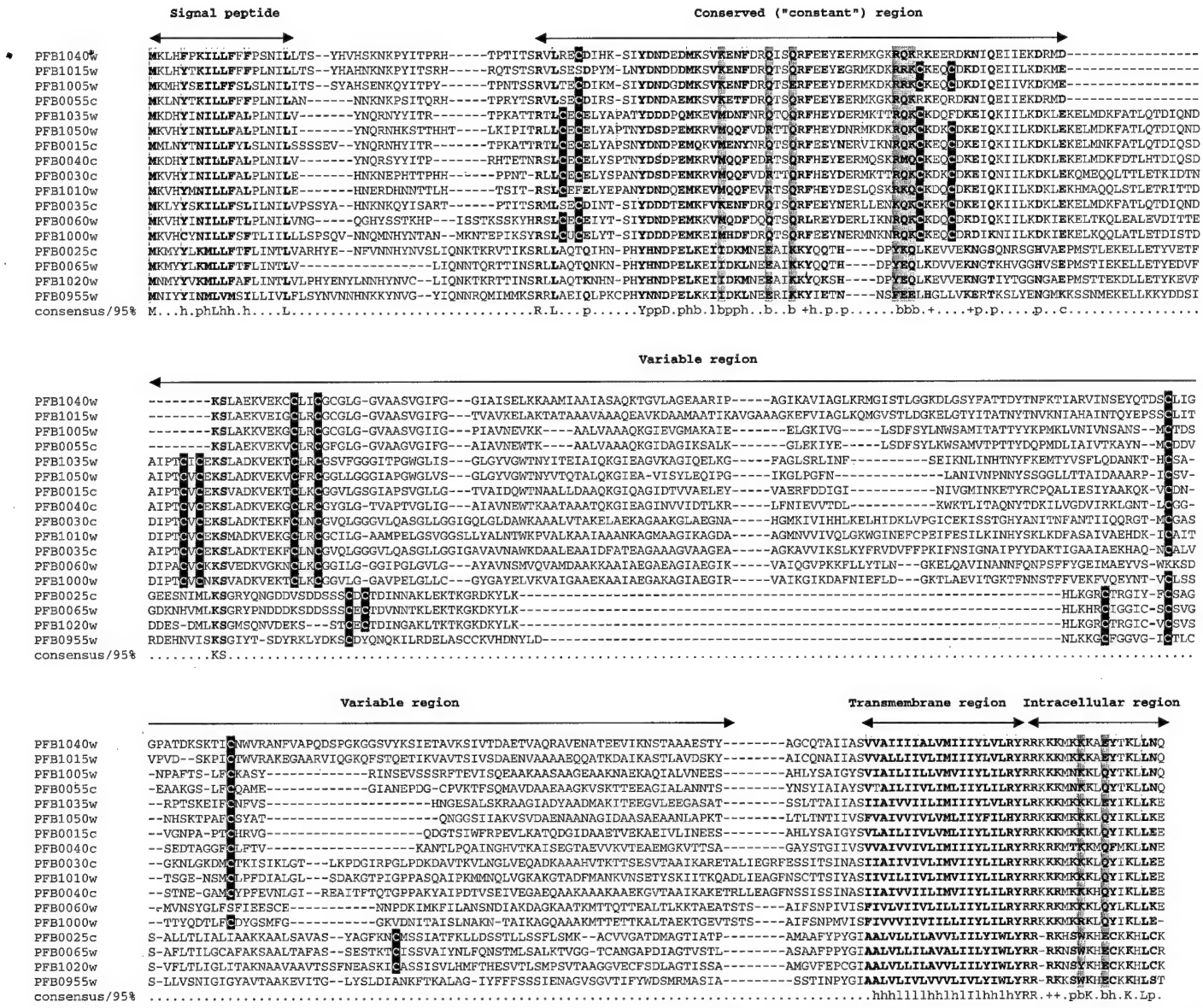


Fig. 3. Multiple sequence alignment of rifins encoded on chromosome 2. The predicted coding regions were aligned with CLUSTALW (*34*) using the default settings. The alignment column shading is based on a 95% consensus, which is shown underneath the alignment; **h** indicates hydrophobic residues (A, C, F, I, L, M, V, W, and Y), **p** indicates polar residues (D, E, H, K, N, Q, R, S, and T), **b** indicates "big" residues (F, I, L, M, V, W, Y, K, R, Q, and E), and + indicates positively charged residues (K and R) (*35*). The cysteines conserved in subsets of rifins are shown by inverse type.

MSP-5, and MSP-1 (a multi-EGF domain protein encoded on chromosome 3) and two *Plasmodium* sexual-stage antigens (*32*) are the only proteins that contain EGF repeats, which suggests that *Plasmodium* obtained the sequence for this domain from its animal host. The plasmodial EGF domains may be involved in parasite adhesion to host cells.

In addition to the families of *Plasmodium*-specific proteins, chromosome 2 contains genes for many secreted and membrane proteins. One of these genes encodes a protein with a modified thrombospondin domain and was transcribed in blood-stage parasites (*17*). Other *Plasmodium* proteins containing thrombospondin domains, such as sporozoite surface protein 2/TRAP and circumsporozoite protein, are involved in the parasitic invasion of host cells (*33*), suggesting that this protein may be involved in the binding of infected red cells to host-cell ligands.

Determination of the first *P. falciparum* chromosome sequence demonstrates that the A+T richness of *P. falciparum* DNA will not prevent the sequencing of the genome. Although technical difficulties not observed during the sequencing of other microbial genomes were encountered, solutions to these problems were found that will facilitate sequencing of the remaining chromosomes. The genome sequence should be of value in the study of *Plasmodium* biology and in the development of new drugs and vaccines for the treatment and prevention of malaria. In addition to these practical benefits, the *Plasmodium* genome sequence should provide broader biological insights, particularly in regard to the plasticity of the eukaryotic genome that is manifest in the preponderance of the predicted nonglobular domains in plasmodial proteins.

References and Notes

1. World Health Organization, *Wkly. Epidemiol. Rec.* **72**, 269 (1997).
2. J. B. Dame *et al.*, *Mol. Biochem. Parasitol.* **79**, 1 (1996); M. Lanzer, D. de Bruin, J. V. Ravetch, *Nature* **361**, 654 (1993); K. Suplick, R. Akella, A. Saul, A. B. Vaidya, *Mol. Biochem. Parasitol.* **30**, 289 (1988); M. J. Gardner, D. H. Williamson, R. J. M. Wilson, *ibid.* **44**, 115 (1991); R. J. M. Wilson *et al.*, *J. Mol. Biol.* **261**, 155 (1996); S. Köhler *et al.*, *Science* **275**, 1485 (1997).
3. R. D. Fleischmann *et al.*, *Science* **269**, 496 (1995).
4. C. M. Fraser *et al.*, *ibid.* **270**, 397 (1995); C. J. Bult *et al.*, *ibid.* **273**, 1058 (1996); C. M. Fraser *et al.*, *Nature* **390**, 580 (1997); J.-F. Tomb *et al.*, *ibid.* **388**, 539

(1997); H. P. Klenk *et al.*, *ibid.* **390**, 364 (1997); C. M. Fraser *et al.*, *Science* **281**, 375 (1998).

5. *P. falciparum* clone 3D7 was selected because it can complete all stages of the life cycle and because 3D7 was used in a genetic cross [D. Walliker *et al.*, *Science* **236**, 1661 (1987)] and in The Wellcome Trust Malaria Genome Mapping Project [J. Foster, J. Thompson, *Parasitol. Today* **11**, 1 (1995)]. Parasites were grown in vitro [W. Trager and W. Jensen, *Nature* **273**, 621 (1978)] and embedded in agarose [D. J. Kemp *et al.*, *ibid.* **315**, 347 (1985)]. Chromosomes were resolved on preparative pulsed-field gels (the process used 1.2% SeaPlaque GTG agarose, a Bio-Rad DRIII apparatus, a 180- to 250-s switch time, a 120° field angle, and 3.7 V/cm for 90 hours at 14°C). Chromosome 2 bands from five gels were adjusted to 0.3 M sodium acetate to prevent melting of the AT-rich DNA and were digested with agarase. The exposure of the DNA to UV light was minimized. A shotgun library of 1- to 2-kb fragments was prepared in pUC18 as described (*3*), except that treatment with *Escherichia coli* DNA polymerase I was performed (0.5 mM deoxynucleoside triphosphates at 37°C for 10 min) after the second ligation step to close nicks before electroporation into DH10B cells. The gel-purified chromosome 2 DNA was only ~85% pure because of the co-migration of sheared DNA from other chromosomes. To compensate for this ~85% purity and to provide excess coverage to compensate for the possible nonrandomness of the shotgun library, we obtained 23,768 sequences (a coverage of about 10-fold). FS+ dye-terminator chemistry (Perkin-Elmer Applied Biosystems, Foster City, CA) was superior to dye-primer chemistry for the sequencing of AT-rich DNA. Sequences were assembled with The Institute for Genomic Research (TIGR) Assembler [G. S. Sutton, O. White, M. D. Adams, A. R. Kerlavage, *Genome Sci. Tech.* **1**, 9 (1995)], which was modified to assemble A+T-rich sequences. Neighboring contigs were identified with the program GROUPER (A. D. Mays, TIGR, Rockville, MD), and 10 groups of 114 contigs were mapped on the chromosome by comparison to sequence-tagged site (STS) markers [M. Lanzer, D. de Bruin, J. V. Ravetch, *Nature* **361**, 654 (1993)]. The closure of physical and sequence gaps was performed as described (*3*). Physical gaps were closed by PCR reactions with a genomic DNA template with primers from adjacent mapped groups or with primers from one mapped group and each of the unmapped groups. PCR reactions (Expand Long Template PCR System, Boehringer Mannheim) contained 100 ng of genomic DNA and 15 pmol of each primer (BioServe Biotechnologies, Laurel, MD) in a 50-ml reaction. Cycling conditions (Perkin-Elmer GeneAmp PCR Systems 9600 or 9700) were as follows: 94°C for 2 min; 10 cycles at 94°C for 1 min, at 50° or 55°C for 1 min, and at 60°C for 2 min; 20 cycles at 94°C for 1 min, at 50° or 55°C for 1 min, and at 60°C for 2 min plus 20 s per cycle; and 1 cycle at 60°C for 10 min. PCR products were purified (QIAquick PCR Purification Kit; QIAGEN, Chatsworth, CA) and sequenced with dye-terminator chemistry. Sequence gaps that were too rich in A+T for primer synthesis and walking were closed by the insertion of the artificial transposon AT-2 [S. E. Devine and J. D. Boeke, *Nucleic Acids Res.* **22**, 3765 (1994)] into the plasmid templates that spanned each sequence gap; multiple transposon-containing subclones of each template were sequenced to close the gaps. The coverage criteria were that every position required at least double-clone coverage (or sequence from a PCR product that was amplified from genomic DNA) and either sequence from both strands or coverage with two different sequencing chemistries. The sequence was edited manually with TIGR Editor, and additional sequencing reactions were performed to improve coverage and to resolve sequence ambiguities. To independently confirm the colinearity of the assembled sequence and genomic DNA, we prepared Nhe I and Bam HI optical restriction maps of chromosome 2 DNA [J. Jing *et al.*, in preparation] and compared them with restriction maps that were predicted from the sequence. The relative errors of predicted and observed fragment sizes were 4.3 and 5.8% for the Nhe I and Bam HI maps, respectively, indicating that the assembled sequence was an accurate representation of the chromosome. Further proof of colinearity was obtained by a comparison of the sequence to a scaffold of YAC-end sequences from chromosome 2 YACs that were isolated from a library provided by K. Hinterberg [J. Foster and J. Thompson, *Parasitol. Today* **11**, 1 (1995); L. Cummings *et al.*, in preparation].
6. D. I. Baruch *et al.*, *Cell* **82**, 77 (1995).
7. Z. Su *et al.*, *ibid.*, p. 89.
8. J. D. Smith *et al.*, *ibid.*, p. 101 (1995); J. A. Rowe, J. M. Moulds, C. I. Newbold, L. H. Miller, *Nature* **388**, 292 (1997).
9. J. L. Weber, *Mol. Biochem. Parasitol.* **29**, 117 (1988).
10. K. D. Vernick and T. F. McCutchan, *ibid.* **28**, 85 (1988).
11. P. Oquendo *et al.*, *ibid.* **18**, 89 (1986); J. Patarapotikul and G. Langsley, *Nucleic Acids Res.* **16**, 4331 (1988).
12. S. Saitoh, K. Takahashi, M. Yanagida, *Cell* **90**, 131 (1997); M. M. Smith *et al.*, *Mol. Cell Biol.* **16**, 1017 (1996); M. M. Mahtani and H. F. Willard, *Genome Res.* **8**, 100 (1998); R. D. Shelby, O. Vafa, K. F. Sullivan, *J. Cell Biol.* **136**, 501 (1997); D. du Sart *et al.*, *Nature Genet.* **16**, 144 (1997).
13. J. Lechner and J. Ortiz, *FEBS Lett.* **389**, 70 (1996); A. A. Hyman and P. K. Sorger, *Annu. Rev. Cell Dev. Biol.* **11**, 471 (1995).
14. The nonredundant (NR) protein sequence database at the National Center for Biotechnology Information (NCBI) (NIH, Bethesda, MD) was searched with the gapped BLAST and PSI-BLAST programs. Coding regions were predicted with GlimmerM, a eukaryotic gene-finding program based on Glimmer [S. L. Salzberg, A. L. Delcher, S. Kasif, O. White, *Nucleic Acids Res.* **26**, 544 (1998)], trained on a set of 117 *P. falciparum* sequences. Gene models based on GlimmerM predictions, similarity of ORFs to known proteins, and prediction of putative signal peptides and transmembrane domains were constructed with ANNOTATOR (L. Xhou, TIGR). In cases where a putative gene had no database match and multiple GlimmerM predictions of gene structure, the highest scoring model was reported. After the first set of models was inspected, it was added to the training set, and GlimmerM was retrained. Gene models should be regarded as preliminary until confirmed by other methods. Protein structural features were delineated with the UniPred program of the SEALS package [D. R. Walker and E. V. Koonin, *Ismb* **5**, 333 (1997)]. Signal peptides were predicted with SignalP [H. Nielsen, J. Engelbrecht, S. Brunack, G. von Heijne, *Protein Eng.* **10**, 1 (1997)], and transmembrane helices were predicted with PHThtm [B. Rost, R. Casadio, P. Fariselli, C. Sander, *Protein Sci.* **4**, 521 (1995)]. Coiled-coil domains were predicted with COILS (J. Kuzio, NCBI). Nonglobular structures were predicted with SEG [J. C. Wooton and S. Federhen, *Methods Enzymol.* **266**, 554 (1996)]. Multiple sequence alignments were constructed with CLUSTALW or with the Gibbs-sampling option of the MACAW program [G. D. Schuler, S. F. Altschul, D. J. Lipman, *Proteins* **9**, 180 (1991); A. F. Neuwald, J. S. Liu, C. E. Lawrence, *Protein Sci.* **4**,1618 (1995)]. Transfer RNAs were identified with tRNAscan [T. M. Lowe and S. R. Eddy, *Nucleic Acids Res.* **25**, 955 (1997)]. Systematic gene names based on a scheme for *S. cerevisiae* [H. W. Mewes *et al.*, *Nature* **387** (suppl.), 7 (1997)] were assigned with the convention PF (for *P. falciparum*), a letter for the chromosome (A for chromosome 1, B for chromosome 2, and so forth), a three-digit code ordering the genes from left to right in increments of five (to allow for the addition of new genes), and a letter denoting the coding strand (w or c, for Watson or Crick strand, respectively).
15. The term "nonglobular" refers to proteins or domains of proteins that do not assume compact, folded structures [J. C. Wootton, *Comput. Chem.* **18**, 269 (1994)]. There is a strong inverse correlation between compositional bias in protein sequences and their ability to fold into a compact, globular domain [J. C. Wootton and S. Federhen, *Methods Enzymol.* **266**, 554 (1996)]. Accordingly, the compositional complexity of a sequence can be used to partition it into predicted globular and nonglobular domains. In this analysis, the prediction was performed with the SEG program with the following parameters: window length, 45; trigger complexity, 3.4; and extension complexity, 3.75.
16. L. Aravind and E. Koonin, unpublished data.
17. D. J. Carucci *et al.*, data not shown.
18. Y. Kim *et al.*, *Nature* **376**, 612 (1995).
19. V. Haucke and G. Schatz, *Trends Cell Biol.* **7**, 103 (1997).
20. A. R. Slabas and T. Fawcett, *Plant Mol. Biol.* **19**, 169 (1992); R. J. M. Wilson, M. J. Gardner, J. E. Feagin, D. H. Williamson, *Parasitol. Today* **7**, 134 (1991).
21. After this manuscript was submitted for publication, we learned of work that confirmed the identification of the 3-ketoacyl-ACP synthase III gene in *Plasmodium* and the importation of nuclear-encoded proteins into the apicoplast in the related parasite *Toxoplasma* [R. F. Waller *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **95**, 12352 (1998)].
22. D. M. Cyr, T. Langer, M. G. Douglas, *Trends Biochem Sci.* **19**, 176 (1994).
23. L. Aravind *et al.*, data not shown.
24. P. Bork, C. Sander, A. Valencia, B. Bukau, *Trends Biochem. Sci.* **17**, 129 (1992); J. Watanabe, *Mol. Biochem. Parasitol.* **88**, 253 (1997); R. L. Coppel *et al.*, *Nature* **310**, 789 (1984); I. A. Quakyi *et al.*, *Infect. Immun.* **57**, 833 (1989); M. Foley, L. Corcoran, L. Tilley, R. Anders, *Exp. Parasitol.* **79**, 340 (1994).
25. S. Bonnefoy, E. Bischoff, M. Guillotte, O. Mercereau-Puijalon, *Mol. Biochem. Parasitol.* **87**, 1 (1997).
26. Sequence data for *P. falciparum* chromosome 3 was obtained from the Sanger Centre (available at http://www.sanger.ac.uk/Projects/P_falciparum/). Sequencing of *P. falciparum* chromosome 3 was accomplished as part of the Malaria Genome Project Consortium with support by the Wellcome Trust.
27. R. R. Hernandez *et al.*, *Mol. Cell Biol.* **17**, 604 (1997); K. Fischer *et al.*, *ibid.*, p. 3679 (1997).
28. B. Knapp, E. Hundt, U. Nau, H. A. Kupper, *Mol. Biochem. Parasitol.* **32**, 73 (1989); B. Knapp, U. Nau, E. Hundt, H. A. Kupper, *ibid.* **44**, 1 (1991); W. B. Li, D. J. Bzik, T. Horii, J. Inselburg, *ibid.* **33**, 13 (1989); B. A. Fox and D. J. Bzik, *ibid.* **68**, 133 (1994).
29. D. G. Higgins, D. J. McConnell, P. M. Sharp, *Nature* **340**, 604 (1989); A. E. Eakin, J. M. Higaki, J. H. McKerrow, C. S. Craik, *ibid.*, **342**, 132 (1989).
30. V. M. Marshall *et al.*, *Infect. Immun.* **65**, 4460 (1997).
31. V. M. Marshall, W. Tieqiao, R. L. Coppel, *Mol. Biochem. Parasitol.* **94**, 13 (1998).
32. L. Aravind, unpublished observations; M. J. Blackman, I. T. Ling, S. C. Nicholls, A. A. Holder, *Mol. Biochem. Parasitol.* **49**, 29 (1991); D. C. Kaslow *et al.*, *Nature* **333**, 74 (1988); P. E. Duffy, P. Pimenta, D. C. Kaslow, *J. Exp. Med.* **177**, 505 (1993).
33. K. J. Robson *et al.*, *Nature* **335**, 79 (1988); C. Cerami *et al.*, *Cell* **70**, 1021 (1992); W. O. Rogers *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **89**, 9176 (1992).
34. J. D. Thompson, D. G. Higgins, T. J. Gibson, *Nucleic Acids Res.* **22**, 4673 (1994).
35. Single-letter abbreviations for the amino acid residues are as follows: A, Ala; C, Cys; D, Asp; E, Glu; F, Phe; G, Gly; H, His; I, Ile; K, Lys; L, Leu; M, Met; N, Asn; P, Pro; Q, Gln; R, Arg; S, Ser; T, Thr; V, Val; W, Trp; and Y, Tyr.
36. We thank the members of the Malaria Genome Sequencing Consortium for the open discussion of data during the development of the effort to sequence the *P. falciparum* genome; D. J. Lipman and L. H. Miller for helpful discussions; M. Gottlieb for support and encouragement; A. Craig for providing the 3D7 clone and for suggestions on pulsed-field gel electrophoresis; P. de la Vega for the culturing of parasites; M. Lanzer for providing STS data; K. Hinterberg for providing the 3D7 YAC library; and the TIGR faculty, sequencing core, bioinformatics staff, and systems administrators for expert advice and assistance. This work was supported by a supplement to the National Institute of Allergy and Infectious Diseases grant R01 AI40125-01, which was made possible with funds from NIH's Office for Research on Minority Health; Department of the Army Cooperative Agreement grant DAMD17-98-2-8005 (to J.C.V.); and Naval Medical Research and Development Command Work Units 61102A.S13.00101.BFX1431, 612787A.870.00101.EFX.1432, 623002A.810.00101.HFX.1433, and STEP C611-102A0101BCX. The opinions and assertions herein are those of the authors and are not to be construed as official or as reflecting the views of the U.S. Navy or Department of the Army.

29 June 1998; accepted 29 September 1998

Sequencing the Malaria Protozoan

By NICHOLAS WADE

Biologists have taken a first step toward cracking the code of the parasite that causes malaria and using the information to explore all the deadly organism's biological weaknesses.

The first of the parasite's 14 chromosomes has been fully sequenced, meaning the order of its DNA units has been determined, by scientists at the Institute for Genomic Research in Rockville, Md., and other institutions.

"This is a very important milestone for malaria research," said Dr. Dyanne Wirth, a malaria expert at the Harvard School of Public Health. "It provides proof of principle that we can sequence the genome and insight into the arrangement of the parasite's genes."

One of the authors of the research, Capt. Stephen L. Hoffman of the Naval Medical Research Center in Bethesda, Md., said that sequencing the genome provided "the foundation for a rational strategy for vaccine and drug development." "What we are generating here is the road map for all malarial research in the 21st century. Without it we are in the Dark Ages," he said. The report appears in the current issue of Science.

Disease experts hope that knowledge of the organism's complete genetic game plan will betray weak points against which new drugs can be targeted. Malaria, once curbed with DDT and chloroquine, is now resurgent in Africa and other parts of the world because the parasite and the mosquitoes that spread it have developed resistance to the usual control measures. Malaria attacks more than 300 million people every year and kills about 2 million, mostly in tropical Africa, according to the World Health Organization.

In the last four years, new techniques of reading the vast stores of information encoded in DNA, the genetic material, have enabled biologists to decipher the biological programming of the smallest organisms, mostly pathogenic bacteria whose genomes, or total DNA, range from one to five million units of DNA.

The malaria parasite, a single-celled animal known as a protozoan, belongs to a higher order of complexity. Its genetic endowment runs to 30 million units of DNA and contains the genes to guide it through the extraordinary transformations of its devious life cycle in the intestine and salivary glands of mosquitoes and in the liver and blood cells of people.

No genome this large has yet been sequenced. But the fact that the protozoan's DNA is packaged into 14 chromosomes, each about the size of a bacterium's genome, encouraged the idea of tackling each one separately. Because of the urgency of developing new defenses against malaria, the National Institutes of Health, the Defense Department and the Wellcome Trust of London formed a consortium to sequence the genome.

> Looking for a genetic soft spot in parasite with a devious life cycle.

The Sanger Center near Cambridge, England, started to sequence chromosomes 1 and 3 while the Institute for Genomic Research, known as TIGR, began with chromosome 2 and has finished its task first.

Dr. J. Craig Venter, until recently TIGR's president, said many experts had predicted that the genome would be impossible to sequence because its DNA does not remain stable when inserted into the bacteria used to copy or clone it. "They even had me spooked," Dr. Venter said.

This problem was overcome by Dr. Hamilton O. Smith, a member of the TIGR team, who discovered that the malarial DNA could be cloned if cut into small enough pieces.

The chromosome, the second shortest in the malarial protozoan's genome, turns out to have 947,103 units of DNA, coding for 210 genes, according to the researchers' report.

Biologists can tell from the nature of the proteins specified by these genes which ones are likely to protrude from the malarial cell, offering possible targets for vaccines, and which ones control metabolic pathways that could be disrupted by drugs.

"Can you imagine having all the pathways laid out?" said Dr. Louis H. Miller, a malarial expert at the National Institutes of Health. "You can do experiments in a day or two that might have taken months."

But a different set of genes is probably switched on at each of the four stages in the protozoan's life cycle. Captain Hoffman said plans were under way to identify the genes that are active at each stage and design countermeasures accordingly.

One of the most unusual features of the malaria protozoan lies in its organelles, units that possess their own DNA and perform special housekeeping duties. Both animal and plant cells possess organelles thought to be ancient bacteria that were long ago enslaved to the cell's needs. Animal cells have mitochondria, which generate energy, and plant cells have plastids that perform photosynthesis.

The malaria protozoan, strangely enough, has both types of organelles. It possesses mitochondria and apicoplasts, organelles that resemble plastids that have lost their equipment for photosynthesis.

Dr. Malcolm J. Gardner of TIGR, a co-author of the new report, noted that the apicoplast's working parts offered particularly good targets for drugs, as they were essentially plant-like and should have no counterparts in the human body.

Protist, Vol. 149, 109–112, May 1998 © Gustav Fischer Verlag

Protist

PROTIST NEWS

The Malaria Genome Sequencing Project

Malaria is caused by apicomplexan parasites of the genus *Plasmodium*. It is a major public health problem in many parts of the world. In 1994 the World Health Organization estimated that there were 300–500 million cases and up to 2.7 million deaths caused by malaria each year, and because of increased parasite resistance to chloroquine and other antimalarials the situation is expected to worsen considerably (WHO 1997). These dire facts have stimulated efforts to develop an international, coordinated strategy for malaria research and control (Butler et al. 1997). Development of new drugs and vaccines against malaria will undoubtedly be an important factor in control of the disease. However, despite recent progress, drug and vaccine development has been a slow and difficult process, hampered by the complex life cycle of the parasite, a limited number of drug and vaccine targets, and our incomplete understanding of parasite biology and host-parasite interactions.

The advent of microbial genomics, i.e. the ability to sequence and study the entire genomes of microbes, should accelerate the process of drug and vaccine development for microbial pathogens. As pointed out by Bloom, the complete genome sequence provides the "sequence of every virulence determinant, every protein antigen, and every drug target" in an organism (Bloom 1995), and establishes an excellent starting point for this process. Today, the complete genome sequences of 13 microbes have been published, including several human pathogens, and many more microbial genomes are in the works (a listing of microbial genome projects completed or underway can be found at www.tigr.org/mdb).

Two main strategies have been used in these projects. One approach, pioneered at TIGR (Fleischmann et al. 1995), is the whole genome shotgun method, in which a genomic library of sheared 1–2 kb fragments is prepared in a plasmid vector, and clones are picked at random and sequenced. Special software is then used to assemble the overlapping fragments into a contiguous sequence. The whole genome method is dependent on high-quality shotgun libraries and robust software for fragment assembly. The second method, used to sequence the *E. coli* genome, for example, involves sequencing of large-insert clones from cosmid or lambda libraries (Blattner et al. 1997). Although not so dependent upon computational resources as the whole genome shotgun method, sequencing of large-insert clones does require a physical map of the genome to guide selection of the clones to be sequenced.

At first, it was unclear how best to proceed in sequencing the genome of *P. falciparum*, the human malaria parasite responsible for the most morbidity and mortality. The *P. falciparum* genome is about 30 Mb in size, about 8- to 10-fold larger than a typical eubacterial genome, and its size was thought to preclude the whole-genome approach due to the computational limitations inherent in the assembly process, and difficulties in closing gaps that usually persist after assembly. The large-insert library approach was ruled out by the fact that *P. falciparum* has an overall base composition of approximately 82% AT. This unusual base composition is thought responsible for the fact that *P. falciparum* DNA is notoriously unstable in *E. coli*, such that representative large-insert (> 20 kb) genomic libraries in plasmid, lambda, and cosmid vectors that could be used for sequencing cannot be prepared. Yeast artificial chromosome (YAC) libraries of *P. falciparum* (Foster and Thompson 1995) have been constructed, however, and while these appear to stably maintain large inserts, YACs are not very well suited to high-throughput sequencing projects.

Abbreviations: EBI: European Bioinformatics Institute; EST: expressed sequence tag; DoD: US Department of Defense; GST: genomic sequence tag; NCBI: National Center for Biotechnology Information; NMRI: Naval Medical Research Institute; PFGE: pulsed field gel electrophoresis; TIGR: The Institute for Genomic Research; TDR: Special Programme for Research and Training in Tropical Diseases; YAC: yeast artificial chromosome.

These problems led to development of a third approach to genome sequencing, namely shotgun sequencing of individual chromosomes purified by pulsed field gel electrophoresis (PFGE). *P. falciparum* has 14 chromosomes ranging from 0.8 to 3.4 Mb in length. Most of the chromosomes of *P. falciparum* clone 3D7 (the clone selected for sequencing) can be resolved in PFGE gels, except for chromosomes 5–9 which co-migrate as a "blob" in the middle of the gel. Chromosomes are resolved on preparative PFGE gels and chromosomal DNA is extracted by agarase digestion. The chromosomal DNA is then sheared into 1–2 kb fragments, cloned into plasmid or M13 vectors, and randomly-picked clones are sequenced. The sequences are assembled to form contigs, and techniques such as PCR from genomic DNA with primers derived from the ends of contigs are used to close gaps in the sequence. Some laboratories also perform limited sequencing of shotgun libraries prepared from YACs previously localized on the chromosomes (Foster and Thompson 1995). The YAC-derived sequences help to group contigs from the same part of the chromosome and assist in gap closure.

Three groups are sequencing the *P. falciparum* genome: TIGR and the Malaria Program of the US Naval Medical Research Institute (NMRI); the Sanger Centre in the UK; and Stanford University. An international consortium including the genome laboratories, bioinformatics centers, and funding agencies was formed to oversee the project, facilitate collaboration, and ensure that the data will be provided to the scientific community in a timely and useful manner (Hoffman et al. 1997). Members of the consortium meet every 6 months to review progress and plan future work. The current status of the project is summarized in Table 1. The strategy of sequencing on a chromosome-by-chromosome basis led naturally to assignment of individual chromosomes to the different genome centers, with the "blob" of currently unresolved chromosomes being undertaken rather heroically by the Sanger Centre. Progress in the first pilot projects, namely chromosome 2 by TIGR/NMRI and chromosome 3 by the Sanger Center, has after initial technical difficulties been good such that both chromosomes are expected to be completed shortly, and the Stanford group has begun work on chromosome 12. Preliminary, unedited data have been released into the public domain and are available for downloading, browsing or searching on web sites maintained at each laboratory (Table 2), the National Center for Biotechnology Information (NCBI), and the European Bioinformatics Institute (EBI). The Sanger Centre and TIGR have started work on the other chromosomes.

Thus despite initial scepticism in the malaria research community that the AT-rich *P. falciparum* genome could be sequenced, the success achieved on chromosomes 2 and 3 proves that it is technically feasible, and malaria researchers should soon have access to the complete genome sequence. Recent technological advances such as stable transfection of *Plasmodium* spp., and microarray technologies for global measurement of gene expression, in combination with the genome sequence, will facilitate research to understand *Plasmodium* biology. In addition, sequencing efforts planned or underway for other *Plasmodium* species and other Apicomplexa such as *Toxoplasma* (Table 2) will provide useful complementary data. Although

Table 1. Chromosome assignments and sequencing status for the Malaria Genome Sequencing Project.

Chromosome(s)[a]	Size (Mb)	Laboratory	Funding[b]	Status (as of 3/98)
1	0.8	Sanger Centre	Wellcome Trust	random sequencing
2	1.0	TIGR/NMRI	NIAID, DoD	annotation
3	1.2	Sanger Centre	Wellcome Trust	closure
4	1.4	Sanger Centre	Wellcome Trust	random sequencing
5–9	1.6–1.8	Sanger Centre	Wellcome Trust	library preparation
10	2.1	TIGR/NMRI	NIAID, DoD	library preparation
11	2.3	TIGR/NMRI	NIAID, DoD	library preparation
12	2.4	Stanford University	BWF	random sequencing
13	3.2	Sanger Centre	Wellcome Trust	library preparation
14	3.4	TIGR/NMRI	BWF, DoD	random sequencing

[a]Estimated sizes for *P. falciparum* clone 3D7 taken from Dame et al. (1996).
[b]NIAID, National Institute for Allergy and Infectious Diseases; DoD, US Department of Defense; BWF, Burroughs Wellcome Fund.

Table 2. Internet resources related to the Malaria Genome Sequencing Project.

Web Site	Content	URL
P. falciparum chromosome 2 TIGR	Preliminary sequence data for chromosome 2.	http://www.tigr.org/tdb/mdb/pfdb/pfdb.html
P. falciparum chromosomes 1, 3, 4 The Sanger Centre	Preliminary sequence data for chromosomes 1, 3, and 4.	http://www.sanger.ac.uk/Projects/P_falciparum/
P. falciparum chromosome 12 Stanford University	Preliminary sequence data for chromosome 12.	http://sequence-www.stanford.edu/group/malaria/index.html
P. falciparum Gene Sequence Tag Project, University of Florida	A collection of ESTs and GSTs for *P. falciparum*.	http://parasite.arf.ufl.edu/malaria.html
Malaria Database Monash Univ., Walter and Eliza Hall Institute	A collection of genetic information on malaria parasites. Sponsored by WHO/TDR.	http://www.wehi.edu.au/MalDB-www/who.html
Malaria Genetics and Genomics National Center for Biotechnology Information (NCBI)	BLAST searches on Apicomplexan sequence data, including *P. falciparum*; *P. falciparum* linkage maps, etc.	http://www.ncbi.nlm.nih.gov/Malaria/
Parasite Genomes Blast Server European Bioinformatics Institute	BLAST searches on sequence data from many parasites, including *Plasmodium.*	http://www.embl-ebi.ac.uk/parasites/parasite_blast_server.html
Malaria Foundation	General information on malaria and many links to malaria-related sites.	http://www.malaria.org/index.htm
Toxoplasma Database University of Pennsylvania	Toxoplasma ESTs clustered with ESTs from dbEST.	http://daphne.humgen.upenn.edu:1024/toxodb/ver_1/
TIGR Microbial Database	A comprehensive listing of microbial genome projects.	http://www.tigr.org/tdb/mdb/mdb.html

it is a long way from laboratory research to the fielding of new drugs or vaccines, with the advent of microbial genomics we can expect the process to be speeded up considerably.

Acknowledgements

The Malaria Genome Sequencing Project is supported by The Wellcome Trust, the US Department of Defense, The Burroughs Wellcome Fund and the National Institutes of Health.This work was supported by a supplement to NIH grant R01-AI40125-01, the Naval Medical Research and Development Command work unit M00101S131720, and NIH-NMRI interagency agreement number Y1AI-6091-01. The opinions and assertions herein are those of the authors and are not to be construed as official or as reflecting the views of the US Navy or naval service at large.

References

Blattner FR, Plunkett Gr, Bloch CA, Perna NT, Burland V, Riley M, Collado VJ, Glasner JD, Rode CK, Mayhew GF, Gregor J, Davis NW, Kirkpatrick HA, Goeden MA, Rose DJ, Mau B and Shao Y (1997) The complete genome sequence of *Escherichia coli* K-12. Science **277**: 1453–74

Bloom BR (1995) A microbial minimalist. Nature **378**: 236

Butler D, Maurice J and O'Brien C (1997) Briefing malaria. Nature **386**: 535–540

Dame JB, Arnot DE, Bourke PF, Chakrabarti D, Christodoulou Z, Coppel RL, Cowman AF, Craig AG, Fischer K, Foster J, Goodman N, Hinterberg K, Holder AA, Holt DC, Kemp DJ, Lanzer M, Lim A, Newbold CI, Ravetch JV, Reddy GR, Rubio J, Schuster SM, Su XZ, Thompson JK and Werner EB (1996) Current status of the *Plasmodium falciparum* genome project. Mol Biochem Parasitol **79**: 1–12

Fleischman RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, Bult CJ, Tomb JF, Dougherty BA, Merrick JM et al. (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. Science **269**: 496–512

Foster J and Thompson J (1995) The *Plasmodium falciparum* genome project: a resource for researchers. Parasitol Today **11**: 1–4

Hoffman SL, Bancroft WH, Gottlieb M, James SL, Bond EC, Stephenson JR and Morgan MJ (1997) Funding for malaria genome sequencing. Nature **387**: 647

WHO (1997) World malaria situation in 1994. Wkly Epidemiol Rec **72**: 269–276

Malcolm J. Gardner[a,1], Hervé Tettelin[a], Daniel J. Carucci[b], Leda M. Cummings[a], Mark D. Adams[a], Hamilton O. Smith[a], J. Craig Venter[a], and Stephen L. Hoffman[b]

[a] The Institute for Genomic Research, 9712 Medical Center Drive, Rockville, MD 20850, USA
[b] Malaria Program, Naval Medical Research Institute, 12300 Washington Avenue, Rockville, MD 20852, USA

[1] Corresponding author;
fax 1-301-838-0208
e-mail gardner@tigr.org

Accession number: txt001dcn; Original accession date: 5 May 1998;
Revision version accession date: 10 June 1998, Archived previous version: yes (txt001dcn5May98)

er molecular medicine

The Malaria Genome Sequencing Project

Daniel J. Carucci, Malcolm J. Gardner, Herve Tettelin, Leda M. Cummings, Hamilton O. Smith, Mark D. Adams, Stephen L. Hoffman and J. Craig Venter

An international consortium of genome centres, advanced development teams and funding agencies has begun the task of sequencing the genome of the parasite *Plasmodium falciparum*, the most important cause of human malaria. Sequencing is proceeding chromosome by chromosome, and the annotated sequence of chromosome 2 is nearly finished. With the continual release of sequence data as they are generated, malaria researchers have access to a steady stream of genomic sequences and will soon have the complete annotation of all of the estimated 5000–7000 *P. falciparum* genes. The task will then be how to best apply these data to the development of new anti-malarial drugs, vaccines and diagnostic tests. This review provides a brief overview of the Malaria Genome Sequencing Project and suggests potential directions for future malaria research.

In 1977, Fred Sanger and his colleagues heralded in the field known as genomics, having shown that it was possible to determine the entire genetic sequence of the virus phi-X (ϕX174) (Ref. 1). In that same year the completed genome of another virus, simian virus 40 (SV40), was also reported (Refs 2, 3). Sequencing progressed rapidly as genomes an order of magnitude larger, such as the bacteriophages T7 and lambda, were completed (Refs 4, 5). Within 15 years of the first viral genome being completed, the genomes of plant chloroplasts and the Epstein–Barr virus (EBV) were finished (Refs 6, 7), and many individual gene sequences were deposited into the public domain (Ref. 8). These first projects, though tedious to complete and requiring a great deal of manual effort, first suggested that 'unlocking the key to the genome' was possible; they also showed that it was technically feasible to determine the entire genome sequence of an organism and thus gain access to the description of its fundamental biology. Automated sequencing apparatus and

Daniel J. Carucci[1] (Corresponding author)
[1]Malaria Program, Naval Medical Research Institute, 12300 Washington Avenue, Rockville, MD 20852, USA, Tel: +1 301 295 6989; Fax: +1 301 295 6171; e-mail: caruccid@nmripo.nmri.nnmc.navy.mil

Malcolm J. Gardner[2], Herve Tettelin[2], Leda M. Cummings[2], Hamilton O. Smith[2], Mark D. Adams[2], Stephen L. Hoffman[1] and J. Craig Venter[2]
[2]The Institute for Genomic Research, 9712 Medical Center Drive, Rockville, MD 20850, USA

Expert Reviews in Molecular Medicine ISSN 1462-3994 http://www-ermm.cbcu.cam.ac.uk

Accession number: txt001dcn; Original accession date: 5 May 1998;
Revision version accession date: 10 June 1998, Archived previous version: yes (txt001dcn5May98)

improvements in sequencing chemistries progressed through the late 1980s (Ref. 9), and soon sequencing laboratories scaled up production, refined computational hardware and software, and developed coordinated methods to produce and analyse large quantities of DNA rapidly and efficiently (Refs 10, 11, 12, 13). Stimulated by the success of previous sequencing projects and the tantalising potential benefits of large-scale sequencing, scientists in the mid-1980s considered the enormous task of determining the sequence of the entire human genome. At 3 billion (3×10^9) base pairs (bp), the Human Genome Project represented the largest genome project ever undertaken. Although most of the initial focus of the Human Genome Project has been the production of genome 'maps', attention is now turning towards sequencing (Ref. 14). Indeed, even before human genome maps were completed, large-scale efforts were directed towards sequencing individual human genes (Refs 10, 11, 12).

Efforts have not been solely centred on the Human Genome Project; the determination of the genetic sequences of microbial organisms, especially those of human pathogens, has the potential to revolutionise the development of new drugs and vaccines. A milestone in genomics was reached in 1995 when the first complete bacterial genome was reported (Ref. 15). The genome of the free-living bacteria *Haemophilus influenzae*, at 1.8 million bp, was the largest completed genome, and the first example of a 'whole-genome sequencing strategy' being applied to a microbial organism. On the heels of this news, Barry Bloom in a leader article in *Nature* considered that 'the power and cost-effectiveness of modern genome sequencing technology mean that the complete genome sequences of 25 of the major bacterial and parasitic pathogens could be available within five years' (Ref. 16). At that time, he thought that 'for about 100 million dollars, we could buy the sequence of every virulence determinant, every protein antigen and every drug target' (Ref. 16). He predicted that this up-front investment in genome sequencing would produce information that would be available to scientists forever, and that 'we could then think about a new, post-genomic era of microbe biology' (Ref. 16).

Genome sequencing is progressing at an extraordinary rate; there are now 13 published microbial genomes and over 60 additional microbial genomes being sequenced (The current status of microbial sequencing can be found at http://www.tigr.org/tdb/mdb/mdb.html). Few people in 1977 could have imagined the advances in genome sequencing that have occurred since the 5386 base pairs of phi-X (ΦX174) were first published. Today, the entire genome of that first sequenced virus could be completed by a handful of people in a genome centre in an afternoon.

Sequencing strategies

The approach taken towards the sequencing of the genome of an organism depends on a variety of factors, including the availability of sequencing reagents, the size of the genome of the organism and other characteristics of the genome. As the size of a given genome project increases, methods are employed to partition the entire genome into smaller subunits. This usually means constructing a DNA library, by dividing the genomic DNA into smaller fragments and cloning them into a vector, such as a cosmid. Cosmids can accept inserts as large as 40 kilobase pairs (kb) and are arranged in order along the genome, creating a physical map of the genome. Once the minimum number of overlapping cosmids has been determined, sequencing begins on each identified cosmid. By sequencing cosmid-sized fragments of the genome, data handling becomes more manageable and assembly of the final sequence is facilitated. However, as the genome size increases beyond several megabase pairs (Mb) (consider, for example, the human genome, which comprises 3×10^9 bp), an additional, upper layer of organisation is needed. The reason for this becomes clear when one considers subcloning the entire human genome using cosmids; a ten-fold representation of the genome (every region in the DNA library should be present at least ten times) would require 750,000 cosmid clones. By today's standards, this would be an unmanageable number of clones to characterise and arrange in order along the genome. For such large organisms, a large-insert library must be created, often using bacterial artificial chromosomes (BACs) (Ref. 17) or yeast artificial chromosomes (YACs; Ref. 18 and reviewed in Ref. 19), which can accept DNA fragments that were several hundred kb in size. A low-resolution physical map (with relatively few clones widely separated) is created using BACs or YACs: these are then subcloned into cosmids to produce a high-resolution map. The

Expert Reviews in Molecular Medicine ISSN 1462-3994 http://www-ermm.cbcu.cam.ac.uk

Accession number: txt001dcn; Original accession date: 5 May 1998;
Revision version accession date: 10 June 1998, Archived previous version: yes (txt001dcn5May98)

production of physical maps requires a tremendous amount of up-front effort in development, characterisation and construction; for example, the generation of a physical map for the Human Genome Project has required nearly 10 years of effort and millions of dollars. In addition, large-insert clones often contain insert DNA that is rearranged (brought together in a different position) or chimeric (two previously separate sections are brought together) and, thus, are of no use for sequencing. Furthermore, DNA in YACs is often contaminated with yeast chromosomal DNA during the purification process. Therefore, new strategies to approach large-genome sequencing are being considered that do not rely on previous mapping data. These include 'random-shotgun sequencing', using small (1–2kb) fragments of sheared DNA, of complete genomes and, for larger genomes, a BAC-end sequencing strategy (Ref. 20). In BAC-end sequencing the minimum number of overlapping clones necessary to cover a region of the genome is identified and sequenced. Only the end of each clone is sequenced; the information is collected and the gaps are then filled by further full-length sequencing. As sequencing technology and computer assembly algorithms improve, it will become possible to sequence larger genomes by the 'shotgun method'. These sequenced-based methods have the potential to expedite genome sequencing and further reduce the overall costs involved.

Sequencing microbial genomes

For smaller genomes, such as those of bacteria that range in size from 600 kb to 5 Mb, shotgun sequencing of the whole genome is becoming routine. The first bacterial genome, *H. influenzae*, was completed almost entirely by 'shotgun sequencing' (Ref. 15); that is, the entire genome (1.8 Mb) was sheared into small (1–3 kb) fragments and randomly cloned into the sequencing plasmid. The development of a coordinated sequencing effort, an integrated database management system, and improved sequence-assembly software meant that the entire genome of *H. influenzae* could be completed with ~24,000 successful sequencing reactions, all within approximately one year. Even as the *H. influenzae* sequence publication 'went to press', sequencing of two additional microbial genomes was already nearing completion (Refs 21, 22). Of the 13 microbial genomes that have been sequenced to date, seven were completed using whole-genome shotgun sequencing.

The Malaria Genome Sequencing Project

In May 1996, at a meeting sponsored by the US National Institutes of Health (NIH) and the Burroughs Wellcome Fund, scientists and funding agencies met to discuss the possibility of sequencing one or more plasmodium genomes, the parasites responsible for human and animal malaria. The result of the meeting was the establishment of an international consortium, comprising genome centres, advanced development teams and funding agencies, whose goal was to sequence and annotate the entire genome of *P. falciparum*, the parasite responsible for nearly all of the deaths due to malaria in humans (Ref. 23). A pilot project at The Institute for Genomic Research (TIGR, Rockville, MD, USA) and the Naval Medical Research Institute (NMRI, Rockville, MD, USA) was funded by the NIH and the US Department of Defense (DoD) to develop sequencing strategies for the Malaria Genome Sequencing Project. This has resulted in the complete 1-Mb sequence of *P. falciparum* chromosome 2 (manuscript, in preparation). Sequencing of the other chromosomes is proceeding at three genome centres: TIGR/NMRI, the Sanger Centre (Hinxton, UK) and Stanford University (Stanford, CA, USA), with funding from the Burroughs Wellcome Fund, the NIH, the Wellcome Trust, and the DoD. Early efforts have met with success and thus the consortium is pushing forward with the intent of completing the entire genome of *P. falciparum* by 2002–2003.

Clinical implications/applications

Why sequence the malaria genome?

The world malaria situation is worsening. The World Health Organization (WHO) estimates that one quarter of the population of the world lives in malarious areas and that 300–500 million cases of malaria occur annually (Ref. 24). Although more than 2.6 million people die every year from this disease, few people in the developed world realise the enormous economic, political and social burden that malaria places on those living with this disease (Ref. 24). In addition, increasing numbers of people will be exposed to malaria as the effects of global warming are manifest and as the mosquito vector of malaria encroaches into the non-malarious world (Refs 25, 26). With increasing air travel in a 'shrinking world', in the

Accession number: txt001dcn; Original accession date: 5 May 1998;
Revision version accession date: 10 June 1998, Archived previous version: yes (txt001dcn5May98)

future, more people that previously were not generally exposed to malaria will be placed at risk from this disease. Unfortunately, drug resistance in *P. falciparum* to chloroquine, one of the best anti-malarial drugs ever developed, is widespread and is found in most of the malarious world. Other species of *Plasmodium*, particularly *P. vivax*, are also beginning to develop patterns of chloroquine resistance (Ref. 27). Success in developing new anti-malarial drugs has been short-lived because plasmodium parasites continue to develop resistance to broad classes of anti-malarial drugs; in fact, most of the drugs used for anti-malarial prevention and therapy such as mefloquine are no longer effective in parts of the malarious world. Moreover, despite numerous clinical trials of malaria vaccines, there is, as yet, no licenced malaria vaccine (Ref. 28). It is clear that novel strategies are urgently needed to combat the menacing problem of malaria.

The difficulty of the situation that faces malaria researchers can be best appreciated when one examines the complexities of the parasite. The malaria parasite is an extraordinarily complex microorganism, which has evolved over the past millennia in a hostile immune environment; there is no apparent symbiosis between the malaria parasite and its human host. The plasmodium parasite possesses a complex multistage life cycle (Fig. 1, fig001dcn) in both a vertebrate (such as human) and an invertebrate (such as an *Anopheles* sp. mosquito) host. It exists (1) free in the circulatory system; (2) inside liver cells (hepatocytes), which are capable of presenting parasite antigens in association with major histocompatibility complex (MHC) molecules; and (3) inside red blood cells, which in humans do not have an MHC-restricted antigen presentation pathway. The parasite is exposed to both humoral (soluble) and cellular immune mechanisms. It has also developed complex drug-resistance mechanisms, which span a broad range of compounds. The design of new anti-malarial drugs and vaccines must, therefore, consider both immune evasion (avoidance of the immune system) and drug-resistance mechanisms. Any new approach must also be directed against multiple stages of the parasite, altogether a momentous undertaking. In many ways, malaria researchers are woefully under-equipped to deal with this complex parasite. Because *in vitro* cultivation of most malaria parasites is routinely possible only for the blood stages, experimental access to the other stages of the parasite life cycle and their respective antigens is limited. Animal models do exist for malaria; however, none reproduces accurately the pathology that is seen in humans. Although the transfection of genes into malaria parasites has been developed recently, it is being used in only a few laboratories and is not yet routine. Finally, of the estimated 5000–7000 genes in *Plasmodium* spp., only a few hundred are known; these represent little more than a brief 'snapshot' of all of the genes used by the parasite. Clearly, more information is needed to develop novel anti-malarial strategies. The advances in genome sequencing over the past two decades now make it possible to consider 'unlocking the malaria genome'. For malaria researchers, access to the malaria genome will undoubtedly provide tools to assist in the discovery of novel targets for the development of malaria vaccines and anti-malaria drugs. It should yield targets for improved diagnostic tests and provide a better understanding of the development of both drug resistance and immune evasion. These should, almost certainly, result in better control of this parasite, and potentially the eradication of malaria in humans.

Genetics and molecular biology

The plasmodium genome

The genome of *Plasmodium* spp. is ~30 Mb and is distributed among 14 chromosomes, which range in size from 650 kb to 3.5 Mb each (Table 1, tab001dcn). Figure 2 (fig002dcn) shows a comparison of the sizes of some other genomes. *Plasmodium falciparum* and several other *Plasmodium* spp. are unusual in that their genomes have an extraordinary bias towards two nucleotides: adenine (A) and thymine (T). In regions that code for proteins, the A-T bias is greater than 76%, whereas in intergenic regions (regions between genes) and in introns (regions within genes that are removed before final transcription), the A-T content can approach 100% (Refs 29, 30). This extreme A-T bias is thought to be responsible for the observed difficulty in cloning and maintaining large segments (greater than several kb) of *P. falciparum* DNA in *Escherichia coli* (Ref. 31). This instability has been problematic because there are, as yet, no bacterial libraries available that can accept large inserts of *P. falciparum* DNA. The development of YACs (Ref. 18) has been applied with success to

Expert Reviews in Molecular Medicine ISSN 1462-3994 http://www-ermm.cbcu.cam.ac.uk

Accession number: txt001dcn; Original accession date: 5 May 1998;
Revision version accession date: 10 June 1998, Archived previous version: yes (txt001dcn5May98)

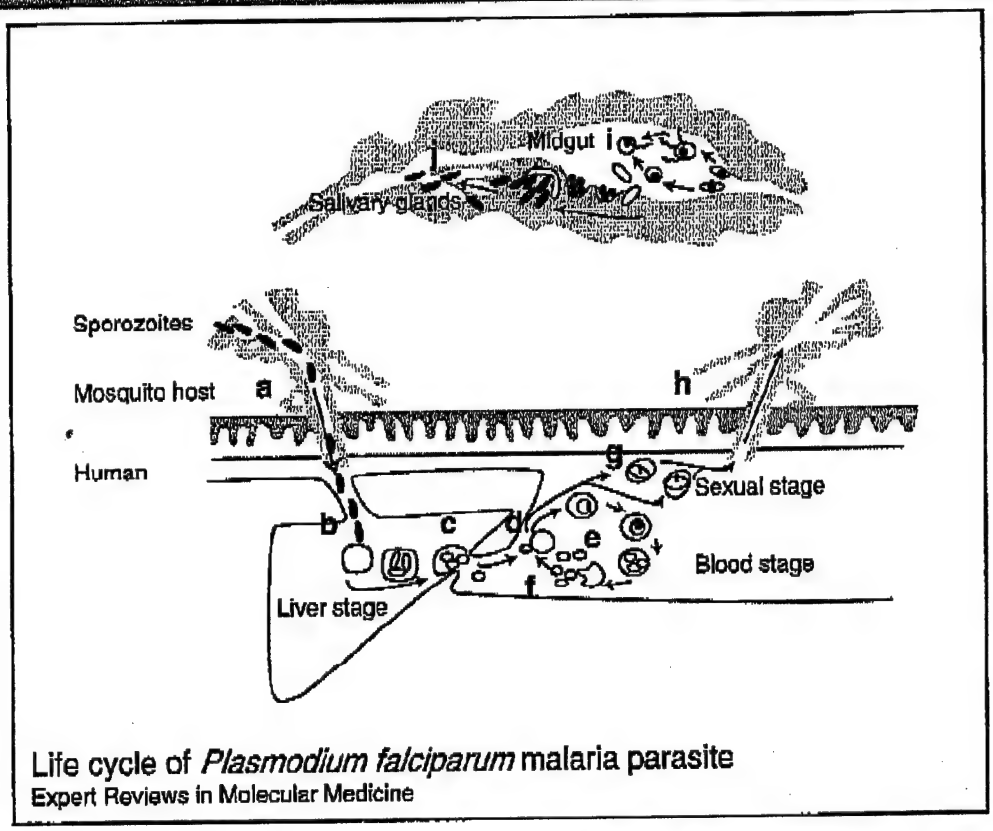


Figure 1. The life cycle of *Plasmodium falciparum* malaria parasite. Malaria is caused by infection with an obligate, intracellular protozoan parasite of the genus *Plasmodium*. Of the four species that infect humans (*Plasmodium falciparum*, *Plasmodium vivax*, *Plasmodium ovale* and *Plasmodium malariae*), it is *P. falciparum* that is responsible for virtually all deaths. The life cycle of *Plasmodium* spp.is complex and somewhat specific to the parasite species. (a) *P. falciparum* infection in humans begins when an infected *Anopheles* sp. mosquito takes a blood meal and injects infective sporozoites into the peripheral circulation. (b) Within minutes, these sporozoites invade hepatocytes in the liver and, over approximately one week, undergo asexual multiplication, producing tens of thousands of merozoite forms of the parasite. (c) When the infected hepatocyte ruptures, merozoites are released into the peripheral circulation. (d) The merozoites invade red blood cells (rbcs) and (e) complete another round of multiplication within 48–72 h, with the production of 16–20 additional merozoites per rbc, which devour the rbc haemoglobin in the process. (f) The released merozoites invade additional rbcs and carry on the cycle. It is the synchronous release of merozoites that is thought to be responsible for the periodic fevers associated with malaria. (g) Some invading merozoites do not divide, but differentiate into male (microgametocyte) and female (macrogametocyte) sexual forms. (h) These sexual forms are taken from the bloodstream by a feeding *Anopheles* sp. mosquito and (i) fertilise in the mosquito midgut to form zygotes. These zygotes further differentiate into motile forms, called ookinetes, migrate through the mosquito gut wall and divide within oocysts on the external gut wall to form thousands of sporozoites. (j) The infective sporozoites are released into the mosquito haemocoele and move to the salivary gland, where they await injection into another human host, thus completing the life cycle **(fig001dcn)**.

P. falciparum (Refs 32, 33) and most recently to *P. vivax* (Ref. 34), presumably owing to the similar nucleotide composition of *Plasmodium* spp. and the yeast *Saccharomyces cerevisiae*. This strategy is being used extensively by malaria researchers (Ref. 35).

Sequencing the *P. falciparum* genome

In designing sequencing strategies for the *P. falciparum* genome project, the consortium focused first on several technical hurdles. The first one was the concern that the high A-T bias and observed inability to produce representative large-insert genomic libraries in *E. coli* would

Accession number: txt001dcn; Original accession date: 5 May 1998;
Revision version accession date: 10 June 1998, Archived previous version: yes (txt001dcn5May98)

Table 1. Sequencing of *Plasmodium falciparum* chromosomes (tab001dcn)

Chromosome	Size (in Mb)	Sequence centre	Status/comment
1	0.65	Sanger	Partially sequenced
2	1.0	TIGR/NMRI	Sequence completed
3	1.2	Sanger	Nearing completion of sequence
4	1.4	Sanger	Partially sequenced
5	1.6	Sanger	
6	1.6	Sanger	
7	1.7	Sanger	
8	1.7	Sanger	
9	1.8	Sanger	
10	2.1	TIGR/NMRI	
11	2.3	TIGR/NMRI	
12	2.4	Stanford	Partially sequenced
13	3.2	Sanger	
14	3.4	TIGR/NMRI	

Abbreviations used and links to further sequence information: Mb = megabase pairs; Sanger = the Sanger Centre (Hinxton, UK), http://www.sanger.ac.uk/Projects/P_falciparum; TIGR = The Institute for Genomic Research (Rockville, MD, USA), http://www.tigr.org/tdb/mdb/pfdb/pfdb.html; NMRI = Naval Medical Research Institute (Rockville, MD, USA), http://www.nmri.nnmc.navy.mil; Stanford = Stanford University (Stanford, CA, USA), http://sequence-www.stanford.edu/group/malaria/index.html

exclude the possibility of using large-insert bacterial libraries to sequence the whole genome using the BAC-end sequencing approach (Ref. 20). The second hurdle was that although YAC libraries of *P. falciparum* were available (Refs 32, 33) YACs were not considered to be good substrates for sequencing; also, bacterial subclone libraries derived from YACs were notorious for their contamination with yeast chromosomal DNA. Finally, the large size of the malaria parasite genome and the intention to divide the sequencing efforts among several laboratories meant that a whole-genome shotgun approach was not practical because sequencing efforts could not be easily partitioned.

Sequencing strategies for *P. falciparum*

The consortium agreed on an approach based on the fact that most of the 14 plasmodium chromosomes can be separated by pulsed-field gradient gel electrophoresis (PFGGE). In fact, using this commonly used molecular biology technique (Ref. 36), over 80% of the genome from *P. falciparum* can be separated as individual chromosomes; the remaining 20% that consists of five co-migrating chromosomes cannot be separated in this way. A decision was made, therefore, to approach the sequencing of the malaria genome one chromosome at a time. The plan was to separate those *P. falciparum* chromosomes that do not co-migrate by PFGGE and treat each one as if it were an individual 1–3-Mb microbial sequencing project. Mapping data that were already available from the ongoing *P. falciparum* Genome Mapping Project, sponsored by the Wellcome Trust (Ref. 35), would provide important information for the process of closing the gaps. Individual chromosomes that were to be sequenced were assigned to three genome centres: TIGR (in conjunction with the NMRI), the Sanger Centre and Stanford University (Table 1, tab001dcn). Random-shotgun libraries that were specific for *P. falciparum* chromosomes were prepared from chromosomes purified using PFGGE and the clones in the libraries were sequenced. In addition, some sequencing centres have used some YAC-based sequencing. At the time of writing, the sequencing of the 1-Mb chromosome 2 (at TIGR/NMRI) and the 1.2-Mb chromosome 3 (at the Sanger Centre) is nearly complete, and significant progress has been made on chromosome 12 (at Stanford University) and on several other chromosomes.

Research in progress and outstanding research questions

The successful sequencing of the first two of the 14 chromosomes of *P. falciparum* has proven the feasibility of completing the entire genome of this parasite. Malaria researchers and genome centres

Expert Reviews in Molecular Medicine ISSN 1462-3994 http://www-ermm.cbcu.cam.ac.uk

Paper number IF 110503

Sequencing the genome of *Plasmodium falciparum*

Daniel J. Carucci[a], Malcolm J. Gardner[b], Hervé Tettelin[b], Leda M. Cummings[b], Hamilton O. Smith[b], Mark D. Adams[b], J. Craig Venter[b] and Stephen L. Hoffman[a]

Advances in microbial genomic sequencing have the potential to revolutionize the control of infectious diseases. Recently, a consortium of researchers and funding agencies from the United States and Great Britain have embarked on a project to sequence the genome from *Plasmodium falciparum*, the most important cause of human malaria. The Malaria Genome Sequencing Project has reached an important milestone with the completion of the entire DNA sequence and annotation of chromosome 2, a 950 kilobase chromosome of *Plasmodium falciparum*. This review article will provide an overview of the malaria genome sequencing project, highlight progress in the field of microbial sequencing, and suggest new directions for future malaria research. Curr Opin Infect Dis 11:000–000. © 1998 Lippincott Williams & Wilkins

[a]Malaria Program, Naval Medical Research Institute, 8901 Wisconsin Avenue, Bethesda, MD 20889, USA; and [b]The Institute for Genomic Research, 9712 Medical Center Drive, Rockville, MD 20850, USA

Correspondence to Dr Daniel J. Carucci at above address.
Fax: +1 301 295 6171; e-mail: carucci@wrlpo.nmri.mil.navy.mil

Current Opinion in Infectious Diseases 1998, 11:000–000

Abbreviation

YAC yeast artificial chromosome

Introduction

Malaria in humans is caused by infection with one of four species of the apicomplexan parasite of the genus *Plasmodium*, *Plasmodium falciparum*, *Plasmodium vivax*, *Plasmodium malariae*, and *Plasmodium ovale* through the bite of a female Anopheles species mosquito. Although it is infection with *P. falciparum* that is the primary cause of human mortality associated with malaria, the other *Plasmodium* species also contribute to the countless episodes of illness in those infected. The WHO has estimated that there are 300–500 million cases of malaria annually and that 1.5–2.7 million people die as a result of malaria each year. The world malaria situation is worsening because of the spread of antimalarial drug resistance, and because of resistance to insecticides by the mosquito vector. In addition, multi-drug-resistant parasites have meant that newly released antimalarial compounds have been routinely short-lived as the malaria parasite has evolved resistance against these new compounds. Indeed, although the concern of drug-resistant malaria has been focused primarily on *P. falciparum*, drug resistance has also extended to *P. vivax*; chloroquine-resistant *P. vivax* is now found in Oceania, southeast Asia, and in parts of South America. A poor understanding of drug resistance mechanisms and a dearth of biochemical drug targets have hindered the development of effective long-lived antimalarial drugs.

To date there is no licensed vaccine against malaria. Evidence from immunological and epidemiological data in malaria endemic areas and from laboratory data in animal and human models, however, suggest that the development of an effective antimalarial vaccine is feasible [1••]. The immunization of mice and humans with radiation-attenuated sporozoites provides 100% protection against malaria [2]. Furthermore, vaccines directed against antigens expressed during the blood stage of the parasite confer protection in mice and non-human primates. The development of novel vaccine delivery systems, particularly DNA vaccines, has the potential to revolutionize vaccinology. DNA vaccines provide a tool for rapidly screening vaccine candidates by circumventing the need to produce recombinant protein immunogens. The development of a successful malaria vaccine will probably require targeting of the numerous antigens expressed at various stages of the complex life cycle of the parasite, and will need to be able to generate the protective immunological responses directed against that antigen. Although the immunological responses against several malaria antigens have been well studied, most malaria researchers would agree that a major

stumbling block to the development of effective antimalarial vaccines is the paucity of well-characterized vaccine targets against particular parasite stages.

The need for new directions

Clearly, the successful development of antimalarial drugs and vaccines will require a more complete understanding of the complex life cycle of the parasite, and the mechanisms underlying drug resistance and the avoidance of protective host immune responses. A detailed study of the genetic code of the parasite may provide the necessary tools for the identification of crucial biochemical targets for the development of antimalarial drugs and reveal targets of protective immunity that might otherwise be undiscovered. Indeed, advances in microbial genomic sequencing have progressed rapidly over the past few years, to a point where it is now possible to determine the entire genetic sequence of an organism. Once deciphered, the entire genetic code of a microbial organism provides researchers with additional tools for the development of novel control and treatment regimens and critical information regarding the biochemistry of the organism. As only a few hundred of the estimated 6000 genes in *P. falciparum* have been identified, and only a handful have been identified as potential targets of vaccines and drugs, the completion of the malaria genome will provide the sequence of every potential drug and vaccine target and will be the foundation for malaria research into the next century.

Introduction to genomic sequencing

The first completed genetic sequences were those of small viruses [3]. These genomes, which were of the order of several thousand base pairs, were deciphered using the then recently developed methods of dideoxynucleic acid sequencing and polyacrylamide gel electrophoresis. Progressively larger genomes were completed over the next 20 years as DNA sequencing technology improved. Automated DNA sequencing, improved chemistries and the computer hardware and software needed to manage the vast amounts of sequence data generated from genome sequencing projects have made the sequencing of large fragments of DNA, including entire genomes, routine. Sequencing centers are depositing hundreds of thousands to millions of bases of DNA sequence each week into public databases. Private companies are also generating vast amounts of DNA sequence data as a means of identifying potential drug targets for the pharmaceutical industry.

Microbial sequencing projects

A landmark in genome sequencing was reached in 1995 with the publication of the first completed genome from a free-living organism, that of *Haemophilus influenzae*. The 1.8 mb genome was the largest such completed genome project and the first to employ a 'whole genome shotgun approach' to an entire genome. Within one year, the genomes of *Mycoplasma genitalium* and *Methanococcus jannaschii* were also published. Recently, the genomes of *Borrelia burgdorferi* [4], the causative agent of Lyme disease; *Helicobacter pylori* [5], implicated in gastric ulcers; and *Mycobacterium tuberculosis* [6], responsible for millions of deaths annually have been completed. Parasite genome projects have also been started, many of which are funded through the WHO/Tropical Disease research program [7,8]. Genome data from microbial pathogens are being generated at an increasing pace, and the implication is that this will result in a better understanding of microbial biology and will lead to more effective antimicrobial therapies. Human pathogens are, however, not the only genomes that are being sequenced. Sequencing is being carried out on genomes from newly discovered families of organisms, such as those that live in extreme conditions or those that have unique metabolic requirements. For example, the genome of *Archaeoglobus fulgidus*, a bacterium that lives at extremely high temperatures and which metabolizes sulfur has recently been completed [9]. The results from these projects and others have shown remarkable differences in how each organism lives. From nearly all completed microbial genome projects between 25 and 35% of each genome codes for proteins that have entirely unknown functions. The success of these projects demonstrates that entire microbial genomes can be elucidated relatively quickly and economically, and have opened the door to additional microbial sequencing projects. There are now over 13 published microbial genomes and more than 60 others in the process of being sequenced. For a current list of microbial genome projects by non-private sources see: http://www.tigr.org/tdb/mdb/mdb.html [10].

Two general approaches have been taken in sequencing these first microbial genomes. The first approach known as 'whole genome shotgun sequencing' involves producing a plasmid library by fragmenting genomic DNA into random, small (1–2 kb pairs) fragments and cloning into a sequencing plasmid. Clones from this library are chosen at random and sufficient numbers are sequenced until an approximately eightfold coverage is achieved. These fragment sequences are then assembled into 'contigs' using specialized computer software forming as near a completed genome as possible. The laborious process of closing gaps between contigs generally requires a combination of methods including additional sequencing, the polymerase chain reaction and others. An advantage of the whole genome shotgun method is that little previous knowledge of the genome is needed. As the computational requirements for assembling the genome are great, this strategy is generally reserved for genomes of the order of a few megabases. For larger genomes, a second approach is traditionally taken. Sequencing centers focus initially on the construction of ordered large insert clone libraries and the construction of a 'physical map' of the genome. A physical map is created by constructing clones (cosmids or lambda vectors) containing large fragments of genomic DNA (generally up to 40 kb pairs) and then determining the minimal overlapping subset of these clones. Each is

then shotgun cloned into sequencing plasmids and randomly sequenced. The genome sequence is generated from both the random shotgun and physical map data. Although a large initial effort is required for the production of the physical map, larger genomes are more easily tackled by this method because the computational requirements of random sequence assembly need be applied only to each large insert clone. As computational hardware and software are improved the random fragment genome sequencing strategy will be applied to larger genomes. A recent effort has been announced that will combine a novel large insert library approach with random shotgun sequencing in order to sequence the entire human genome (3 billion base pairs) in 3 years [11].

Sequencing the malaria genome

An international consortium of funding agencies and genome centers was formed in 1996, whose goal was the completion of the entire genomic sequence of *P. falciparum* [12]. Several important technical hurdles were addressed, including the observed instability of large fragments of *P. falciparum* DNA in typical bacterial plasmid systems, the large genome size, and the means to partition the sequencing efforts among the members of the consortium [13]. The greatest perceived challenge was the fact that there had been little success in the production of a bacterial plasmid or lambda vector library containing inserts more than several kilobases in size, and that plasmids containing even small inserts were often rearranged. The extreme nucleotide composition of *P. falciparum* and other *Plasmodium* species with the percentage of A and T in coding regions reaching 76% and in non-coding regions approaching 100% was thought to be responsible for this instability in bacteria in which the AT composition is of the order of 50%. This instability has not, however, been problematical in yeast systems, in fact stable large insert libraries have been constructed using yeast artificial chromosomes (YAC), and these libraries have been used extensively in a Wellcome Trust-funded *P. falciparum* genome mapping project. Unfortunately, YAC clones are generally considered to be poor substrates for sequencing, and random shotgun libraries derived from purified YAC clones are notorious for yeast chromosomal DNA contamination. Pilot projects were initiated to address this instability; however in the short term the use of a large insert library approach to the sequencing project was considered impractical, although the limited sequencing of YAC clones was considered to be potentially useful.

The genome of *P. falciparum* is large, approximately 30 mb in size, and is distributed on 14 chromosomes, which range in size from 650 kb to 3.4 mb, and two extrachromosomal elements, a 35 kb plastid-like element and a 6 kb mitochondrial repeat element. As approximately 80% of the genome can be resolved by pulsed field gel electrophoresis, the consortium agreed that the genome should be tackled chromosome by chromosome using the random shotgun sequencing approach. This strategy facilitated sequencing of the genome by distributing the entire genome effort among several sequencing centers, with each chromosome project equivalent to a 1–3 mb microbial sequencing project. Sequencing of the *P. falciparum* genome is being carried out by The Institute for Genomic Research and Naval Medical Research Institute (USA), The Sanger Center (UK) and Stanford University (USA) (see Table 1). A pilot project by The Institute for Genomic Research and the Naval Medical Research Institute was funded by the National Institute of Allergy and Infectious Diseases and the US Department of Defense to determine the sequence of chromosome 2 from *P. falciparum* (1 mb) and has been completed (Gardner MJ *et al.*, submitted). At the time of writing, the sequence of chromosome 3 (1.2 mb) is nearing completion by the Sanger Centre UK with funding by the Wellcome Trust, and the chromosome 12 project is well underway at Stanford University with funding by the Burroughs Wellcome Fund. Projects to complete the remaining 11 chromosomes have been started by these three sequencing centers, and the sequence data being generated is released into the public domain on a regular basis. The completion of these first chromosomes has validated the approach adopted by the consortium and has given credence to the hope that the entire genome will soon be completed.

Conclusion

A critical goal for the 'next genome project' will be the use of genomic sequence data derived from the Malaria Genome Project towards the development of novel antimalarial drugs and vaccines, as well as establishing a more complete understanding of parasite biology, host interactions, and immune evasion and drug resistance mechanisms. The challenge is to bring the vast amount of genome sequence data from the genome project to

Table 1. Distribution of *P. falciparum* chromosome sequencing

Chromosome	Genome center
1	Sanger Centre
2	TIGR/NMRI
3	Sanger Centre
4	Sanger Centre
5	Sanger Centre
6	Sanger Centre
7	Sanger Centre
8	Sanger Centre
9	Sanger Centre
10	TIGR/NMRI
11	TIGR/NMRI
12	Stanford University
13	Sanger Centre
14	TIGR/NMRI

TIGR/NMRI, The Institute for Genomic Research/Naval and Medical Research Institute

bear on malaria research. Technologies such as DNA microarrays are already being developed to explore gene regulation and expression on a genome-wide scale in humans and microbes [14•,15,16]. Using these and other technologies it will be possible to study the stage-specific expression of the entire *P. falciparum* genome. By studying gene expression from various differing phenotypic isolates, for example those that differ in their resistance to a particular drug, a more precise study of the mechanisms of drug resistance will be possible. For vaccine development, gene expression studies will be useful for identifying potential targets expressed in blood-stage parasites, and for drug development they may be particularly useful for identifying the expression of novel biochemical targets. As a result of current technical limitations, it may be difficult to apply these technologies to the study of gene expression in some life cycle stages thought to be critical for vaccine development, i.e. the infected hepatocyte. Furthermore, gene expression studies alone may provide little information regarding the subcellular localization of the translated protein and may not be well correlated with protein expression, especially secreted proteins [17]. Other strategies will be needed to identify, characterize and validate the thousands of vaccine targets that will be identified from the genome project. A 'big science' approach comparable to the magnitude of the genome project will be required. For vaccine design, an approach has been proposed to produce DNA vaccines against each individual open reading frame, and by immunizing groups of mice to produce antibodies against the encoded protein [2]. These antibodies can then be used to screen protein expression at each stage of the parasite by immunofluorescence testing and then characterize protein expression and location. A subset of vaccine candidates are then selected on the basis of the pattern of protein expression at the desired stage of the life cycle. Ultimately, the selection of optimal vaccine and drug candidates and the development of an optimal antimalarial vaccine and drug will require a combination of computer modelling, bioinformatics, gene expression studies and protein expression analyses, but the completion of the entire genome of *P. falciparum* will provide the foundation for all these studies and give hope that control of this devastating disease is within reach.

Acknowledgements

The opinions and assertions herein are those of the authors, and are not to be construed as official or as reflecting the views of the US Navy or naval service at large. The work was supported by the Office for Research on Minority Health of the National Institutes of Health and by the Naval Medical Research and Development Command work units STO F 6.161102AA0101BFX and STEP C611102A0101BCX.

References and recommended reading

Papers of particular interest, published within the annual period of review, have been highlighted as:
- • of special interest
- •• of outstanding interest

1 Doolan DL, Hoffman SL. Multi-gene vaccination against malaria: a •• multistage, multi-immune response approach. Parasitol Today 1997; 13:171–178.
This review explores the current malaria vaccine development and highlights strategies to target the complex life cycle of the parasite using novel strategies, including the use of DNA vaccines.

2 Hoffman SL, Carucci DJ, Rogers WD. Using DNA-based vaccine technology and the malaria genome project to overcome obstacles to malaria vaccine development. In: Malaria: parasite biology, pathogenesis, and protection. Sherman IW (editor): Washington, DC: ASM Press; 1998; [pp. 545–556.

3 Watson JD. The human genome project: past, present, and future. Science 1990; 248:44–49.

4 Fraser CM, Casjens S, Huang WM, Sutton GG, Clayton R, Lathigra R, et al. Genomic sequence of a Lyme disease spirochaete, *Borrelia burgdorferi*. Nature 1997; 390:580–586.

5 Tomb JF, White O, Kerlavage AR, Clayton RA, Sutton GG, Fleischmann RD, et al. The complete genome sequence of the gastric pathogen *Helicobacter pylori*. Nature 1997; 388:539–547.

6 Cole ST, Brosch R, Parkhill J, Garnier T, Churcher C, Harris D, et al. Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. Nature 1998; 393:537–544.

7 Tanaka M, Tanaka T, Inasawa J, Nagatsuchi S, Mutsui Y, Kaukas A, et al. Proceedings of the Schistosome Genome Project. Mem Inst Oswaldo Cruz 1997; 92:829–836.

8 Santos MRM, Cano MI, Schijman A, Lorenzi H, Vazquez M, Levin MJ, et al. The Trypanosoma cruzi genome project: nuclear karyotype and gene mapping of clone CL Brener. Mem Inst Oswaldo Cruz 1997; 92:821–828.

9 Klenk HP, Clayton RA, Tomb JF, White O, Nelson KE, Ketchum KA, et al. The complete genome sequence of the hyperthermophilic, sulphate-reducing archaeon *Archaeoglobus fulgidus*. Nature 1997; 390:364–370.

10 The Institute for Genome Research. [web page] http://www.tigr.org/tdb/mdb/mdb.html [26 August 1998].

11 Venter JC, Adams MD, Sutton GG, Kerlavage AR, Smith HO, Hunkapiller M. Shotgun sequencing of the human genome. Science 1998; 280:1540–1542.

12 Hoffman SL, Bancroft WH, Gottlieb M, James SL, Burroughs EC, Stephenson JR, Morgan MJ. Funding for the malaria genome sequencing project. Nature 1997; 387:647.

13 Gardner MJ, Tettelin H, Carucci DJ, Cummings LM, Adams MD, Smith HO, et al. The Malaria Genome Sequencing Project. Protist 1998; 149:109–112.

14 DeRisi JL, Iyer VR, Brown PO. Exploring the metabolic and genetic control • of gene expression on a genomic scale. Science 1997; 278:680–686.
This article reports the results of a comprehensive analysis of essentially every gene from the yeast, *S. cerevisiae*, accompanying the metabolic shift from fermentation to respiration. It highlights the power of DNA microarrays to examine thousands of genes simultaneously.

15 Lashkari DA, DeRisi JL, McCusker JH, Namath AF, Gentile C, Hwang SY, et al. Yeast microarrays for genome wide parallel genetic and gene expression analysis. Proc Natl Acad Sci U S A 1997; 94:13057–13062.

16 National Institutes of Health, National Cancer Institute Cancer Genome Anatomy Project [web page] http://www.ncbi.nlm.nih.gov/ncicgap/expression_tech_info.html. [26 August 1998].

17 Anderson L, Seilhamer J. A comparison of selected mRNA and protein abundances in human liver. Electrophoresis 1997; 18:533–537.

DRAFT

Optical Mapping of *Plasmodium falciparum* chromosome 2.

Junping Jing[1], Christopher Aston[1], Zhongwu Lai[1], Daniel J. Carucci[2], Malcolm J. Gardner[3], J. Craig Venter[3], David C. Schwartz[1,4]

[1]W. M. Keck Laboratory for Biomolecular Imaging
New York University
Department of Chemistry
31 Washington Place
New York, New York 10003

[2]Malaria Program, Naval Medical Research Institute
12300 Washington Ave, Rockville, MD 20852

[3]The Institute for Genomic Research
9712 Medical Center Drive
Rockville, MD 20850

[4]Corresponding author: David C. Schwartz, Ph.D.
W. M. Keck Laboratory for Biomolecular Imaging
Department of Chemistry, Room 866
New York University, 31 Washington Place
New York, New York 10003
PHONE 212-998-8429, FAX 212-995-8487
E-Mail schwad01@mcrcr.med.nyu.edu

ABSTRACT

Detailed restriction maps of microbial genomes are a valuable resource in genomic sequencing studies but are toilsome to construct by combining maps derived from cloned DNA. Analysis of genomic DNA enables large stretches of the genome to be mapped and circumvents library construction and associated cloning artifacts. We used pulsed field gel-purified *P. falciparum* chromosome 2 DNA as the starting material for Optical Mapping, an approach for making ordered restriction maps from ensembles of single DNA molecules. DNA molecules were bound to derivatized glass surfaces, cleaved with *Nhe* I or *Bam*H I and imaged by fluorescence microscopy. Large pieces of the chromosome containing many DNA fragments were mapped. Maps were assembled from 50-60 molecules generating an average contig depth of 15 molecules and a high resolution consensus restriction map was generated.The maps were used as a means to verify assemblies from the plasmid library used for sequencing. Maps generated *in silico* from the sequence data also corresponded with the optical maps. Such high resolution restriction maps may become an indispensable resource for large-scale genome sequencing projects.

INTRODUCTION

Optical mapping is a system for the construction of ordered restriction maps from single molecules (Schwartz et al. 1993, Samad et al. 1995). Individual DNA molecules bound to derivatized glass surfaces, which have been cleaved with restriction enzymes, are imaged by fluorescence microscopy. Cut sites are visualized as gaps between cleaved DNA fragments which retain their original order (Cai et al. 1995, Cai et al. 1998). Optical Mapping has been used to prepare maps of a number of large insert clone types such as bacterial artificial chromosomes (Cai et al. 1998) and most recently genomic DNA (Lin et al. 1998). Large fragments of randomly sheared DNA are mapped with high cutting efficiency and the many overlapping restriction site landmarks allow contigs to be assembled. A shotgun mapping strategy can thus be employed. Optical mapping of genomic DNA enables large stretches of a genome to be mapped, which simplifies contig

formation. Library construction is obviated enabling mapping of organisms such as *Plasmodium falciparum* (*P. falciparum*) with AT-rich DNA, which is difficult to clone. Also, cloning artifacts are precluded enabling more accurate maps to be generated. Furthermore, small amounts of starting material are required facilitating the mapping of parasites, which are problematic to culture. Such restriction maps provide a picture of the architecture of large spans of the genome and have value in shotgun sequencing. They provide an ideal scaffold for sequence assembly, finishing and verification. Gaps between contigs can be characterized in terms of location and breadth, thereby facilitating closure techniques.

Sequencing of chromosome 2 of *P. falciparum* was recently completed by Gardner and colleagues (Gardner et al. 1998), as part of an international consortium sequencing the whole *P. falciparum* genome (Foster 1995, Dame et al. 1996). Existing physical maps of *P. falciparum* chromosomes (chromosome 3 [Thompson and Cowman, 1997] and chromosome 4; [Watanabe and Inselberg, 1994, Sinnis and Wellems, 1988]), prepared by restriction digestion, gel fingerprinting and hybridization of probes are low resolution and not suitable for sequence verification. In order to investigate the feasibility of optical mapping of a whole eukaryotic chromosome we constructed high resolution, ordered restriction maps *P. falciparum* chromosome 2 using genomic DNA and later compared these maps with those generated *in silico* from the sequence data.

MATERIALS AND METHODS

Parasite preparation

P. falciparum (clone 3D7) was cultivated using standard techniques (Trager and Jensen, 1976). In order to minimize possible alterations of the genome that can occur in continuous culture (Corcoran et al. 1986), parasite aliquots were kept frozen in liquid N_2 until needed and then cultivated only as long as necessary. Parasites were cultivated to late trophozoite/early schizont stages and enriched on a Plasmagel gradient. The parasitized red blood cells were washed once with several volumes of 10 mM Tris (pH 8), 0.85% NaCl and the parasites were freed from the erythrocytes by incubation in ice cold 0.5% acetic acid in dH_2O for 5 min, followed by several washes in cold buffer. The parasites were resuspended to a concentration of 2×10^9/ml in buffer and maintained in a 50°C waterbath. An equal volume of 1% InCert agarose in buffer, prewarmed to 50°C, was mixed with the prewarmed parasites and the mixture was added to a 1 cm × 1 cm × 10 cm gel mold, plugged at one end with solidified Insert agarose, and was allowed to cool to 4°C. The agarose-embedded parasites were pushed out of the mold and incubated with 50 ml proteinase K solution (2 mg/ml proteinase K in 1% sarkosyl, 0.5 M EDTA) at 50°C for 48 hrs with one change of proteinase K solution and were stored in 50 mM EDTA at 4°C.

Chromosome 2 isolation by pulsed field gel electrophoresis (PFGE)

Uniform parasite slices were taken with a glass coverslip using two offset microscope slides as guides. One half to one quarter of a single slice was sufficient per lane. Parasite slices were arranged end to end on the flat side of the gel comb. The parasites were fixed to the comb by a small bead of molten (60°C) agarose. The comb was then placed into the gel mold and molten agarose (1.2% SeaPlaque[FMC, Rockland, ME] in 0.5× TBE) poured around the parasite-containing slices. Once cooled, the comb was removed and the space filled with molten agarose. A CHEF DRIII apparatus (Bio-Rad, Hercules ,CA) was used for all chromosome separations. Gels were run with 180-250 sec ramped pulse time at 3.7 V/cm and 120° field angle, for 90 hrs at 14°C with

recirculating buffer at approximately 1 l/min, using *Saccharomyces cerevisiae* and/or *Hansenula wingei* CHEF size markers (Bio-Rad). To minimize UV damage to the DNA, gel slices were removed from the ends of the gel, stained with ethidium bromide (5μg/ml) and visualized by long wave (320 nm) UV light. Notches corresponding to the individual chromosomes were made in the agarose gel and used as guides to cut the chromosome from the gel. The chromosome-containing gel slices were stored in 50 mM EDTA at 4°C until needed. The gel was stained with ethidium bromide to verify the chromosome excision. The genome of *P. falciparum* is approximately 30 Mb in size, consisting of 14 chromosomes ranging in size from 0.6-3.5 Mb (Foote and Kemp 1989). PFGE resolves most of the *P. falciparum* chromosomes, except 5-9 which are similar sizes and co-migrate. The gel band containing *Plasmodium falciparum* chromosome 2 was easily resolved, cut from the gel, melted at 72°C for seven minutes and incubated with agarase at 40°C for two hours. The melted agarose band was diluted in TE to a final DNA concentration suitable for optical mapping (~20 pg/μl).

Mounting and digestion of DNA on optical mapping surface

Optical mapping surfaces were prepared as previously described (Aston et al. 1998). Briefly, glass coverslips (18 × 18 mm^2; FISHERfinest, Pittsburgh, PA) were cleaned by boiling in concentrated nitric, then hydrochloric acid. Surfaces were derivatized with 3-aminopropyldiethoxymethyl silane (APDEMS; Aldrich Chemical, Milwaukee, WI). One surface was placed onto a microscope slide. 10 μl of DNA sample was added to the edge between the surface and the slide and spread into the space between the surface and the slide. The surface was then peeled off from the slide. Digestion was performed by adding 100 μl of digestion solution (50 mM NaCl, 10 mM Tris·HCl [pH 7.9], 10 mM $MgCl_2$, 0.02% Triton-x 100, 20 units of restriction endonuclease;New England Biolabs, Beverly, MA) onto the surface and incubating at 37°C from 15 min to 30 min. The buffer was aspirated and the surface washed with water before staining of DNA with YOYO-1 (Molecular Probes, Eugene, OR) homodimer, prior to fluorescence microscopy. Co-mounted lambda bacteriophage DNA was used as a sizing standard and also to estimate cutting efficiencies.

Image acquisition, processing and map construction

DNA molecules were viewed by fluorescence microscopy. The Optical Mapping surface was scanned by the operator for individual digested DNA molecules of adequate length and quality to be collected for image processing and map making. Images were collected with a charge coupled device (CCD) camera (Princeton Instruments, Trenton, NJ) using Optical Map Maker (OMM) software, as previously described (Jing et al. 1998). Images of DNA fragments were processed using a modified version of NIH Image (Huff, Ph.D. thesis) which integrates fluorescence intensity for each fragment. These values were used to assemble an ordered restriction map for each molecule. Maps were manually constructed using Microsoft Excel spreadsheets. Fluorescence intensity of lambda bacteriophage DNA standards was used to measure the size of the *P. falciparum* restriction fragments on a per image basis. Cutting efficiences (on a per image basis) were determined from scoring cut sites on sizing standard molecules contained in the same field as the genomic DNA molecules. Standard molecules were cut once by *Nhe* I and five times by *Bam*H I. The map for the entire chromosome 2 was manuallyassembled into contigs by aligning overlapping regions of congruent cut sites. If there were no overlapping regions, the molecules were considered to be from a contaminating *P. falciparum* chromosome and were discarded. Consensus maps for chromosome 2 were assembled by averaging the fragment sizes from the individual maps derived from maps underlying the contigs.

Southern blotting of *P. falciparum* genomic DNA

10 μg *P. falciparum* genomic DNA was digested with *Nhe* I or *Bam*H I, resolved by PFGE (POE apparatus, box length 20 cm, 1% gel in 0.5× TBE, pulse time 1 sec; 2 sec, switch time 12 sec, 150 volts, for 24 h) (Schwartz and Koval, 1989), blotted and hybridized with probes derived from small insert clones used for sequencing (PF2CM93 and PF2NA66). Probes were labeled by random priming.

RESULTS

***P. falciparum* chromosome 2 DNA sample**

A chromosome 2 gel slice was used as starting material. Despite the AT-rich nature of the *P. falciparum* genome (80-85%), melting of low gelling temperature agarose inserts did not affect the integrity of the DNA and the chromosome was competent for optical mapping.

Improved DNA mounting technique

Previously, we mounted DNA sample onto optical mapping surface by sandwiching the sample between an optical mapping surface and a microscope slide and then peeling the surface from the slide. DNA molecules were stretched and fixed onto the surface. This method works very well with bacteriophages, cosmids and BACs (Cai et al. 1995, Cai et al. 1998); however, larger genomic DNA molecules tend to form crossed molecules. We improved this approach by adding the sample to the edge formed by placing a surface onto a slide. The liquid DNA sample spread into the space between the surface and the slide by capillary action. Consequently, DNA breakage was minimized, molecules tended be stretch in the same direction, and crossed molecules were also minimized (see Fig. 1).

***Nhe* I and *Bam*H I maps for *P. falciparum* chromosome 2**

The genomic DNA was mapped with either *Nhe* I (Fig. 1A) or *Bam*H I (Fig. 1B). Fragment sizes were calculated by comparison with co-mounted lambda bacteriophage DNA (48.5 kb). *P. falciparum* DNA has an AT content of 80-85% and lambda bacteriophage DNA has an AT content of 50%. The YOYO-1 fluorochrome used for DNA staining preferentially intercalates between GC pairs. A correction factor was therefore applied to each fragment size to correct for massively different fluorochrome incorporation (Netzel et al. 1995). Lambda bacteriophage DNA was also used to determine areas on the surface where cutting efficiency was highest. Cutting efficiencies were in excess of 80%. Maps were obtained from individual molecules of about 350 kb. Consensus maps were assembled from 50-60 molecules generating an average contig depth of 15 molecules. Chromosome 2 was found to be 976 kb by optical mapping with *Nhe* I and 946 kb by optical mapping with *Bam*H I (average size 961 kb). There were 40 fragments in the *Nhe* I map, ranging from 1.5 kb-115 kb, with average fragment size 24

kb (Fig. 2A). There were 30 fragments in the *Bam*H I map ranging from 0.5 kb-80 kb, with average fragment size 32 kb (Fig. 2B). Each fragment size in the consensus map was averaged from 10-15 fragments. Although *P. falciparum* chromosome 2 migrates as a distinct band by PFGE, we found the gel slice to contain only 60% chromosome 2-specific DNA. The remaining optical mapping data was rejected.

Integration of optical maps and sequence data

The chromosome 2 sequence assembled by Gardner and colleagues (Gardner et al. 1998) is a large contig covering most of the chromosome. The optical restriction maps were compared to restriction maps predicted from the sequence, and there was very good correspondence between the two, indicating that there were no major rearrangements or errors in the assembled sequence (Table 1). The optical map included all fragments above 500 bp predicted from sequence. The overall agreement between these maps and the sequence was therefore excellent, with the average fragment size difference below 600 bp (relative error 4.3%) for the *Nhe* I map. The average fragment size difference for the *Bam*H I map was 1.2 kb (relative error 5.8%). However, there were several notable differences. Large differences in size for the fragments at each end of the chromosome were noted fragments (Tables 1 and 2). This is because the sequence for these subtelomeric regions is still under construction. PCR products spanning subtelomeric gaps are currently being sequenced. The optical map sizes were larger than those predicted from sequence for certain other fragments (Tables 1 and 2). These differences were due to falsely large fluorescence intensity measurements caused by crossed molecules. Currently, we integrate length measurements with fluorescence intensity measurements to improve on our sizing of these fragments. Chromosome 2 maps using these new measurements show no exceptional errors (not shown; work in progress). The map was used to facilitate sequence verification. Optical maps can also be used at the earlier sequence assembly stage to form a scaffold for assembly of contigs formed from sequencing.

Map confirmation by southern blotting

In order to confirm the optical maps, pulsed field gels of total *P. falciparum* DNA digested with *Nhe* I or *Bam*H I were generated. Plasmid clones used as sequencing templates were used as probes on southern blots of the gels. Restriction fragment sizes of the blots were closely comparable in size to the fragments seen on the optical maps and those predicted from the preliminary sequence. Probe PF2CM93 hybridized to a 7.5 kb band generated by *Nhe* I digestion and PFGE. The fragment size predicted from sequence information was 7.6 kb. The corresponding fragment size from the optical map was also 7.6 kb (Table 1). The same probe hybridized to a 41 kb band generated by *Bam*H I digestion and PFGE. The fragment size predicted from sequence information was 41.3 kb. The corresponding fragment size from the optical map was 40.8 kb (Table 2). Probe PF2NA66 also generated data with fragment sizes that were very similar (Tables 1 and 2). By using the same probe on DNA digested with the two different enzymes, the optical maps were oriented and linked with one another.

DISCUSSION

We have generated a high resolution optical restriction map of *P. falciparum* chromosome 2 which was used sequence verification. The maps can also be used for final sequence assembly (Cai et al. 1998). The fidelity of the optical maps was checked by using southern blotting. Firstly, this enabled the optical maps to be cross checked against the sequence. This approach could be useful for assembling data acquired using different techniques and would allow the placement of very short sequence contigs onto a map. Secondly, the two maps were also oriented and linked relative to each other. Linking of single-enzyme maps produces a much higher resolution multi-enzyme map which is rich in information. Smaller contigs can be placed on a multienzyme map (Cai et al. 1998). This approach could also be used to assign STS markers or ESTs to restriction fragments on a whole genome optical map.

Despite the fact chromosome 2 is easily resolved by PFGE, we found the chromosome 2 gel slice to contain only 60% chromosome 2-specific DNA. Consequently, a lot of the optical mapping data was rejected. Should we have mapped other chromosomes using the same strategy we could not predict the acquisition of clean

data from chromosomes which are less resolvable by PFGE, such as chromosomes 5-9. Current optical mapping studies on *P. falciparum* use whole genomic DNA as starting material. The chromosomes are resolved at the level of data rather than as physical entities. The data segregates into 14 deep contigs corresponding to the various chromosomes. Chromosome 2 can be resolved based on size and the near complete correspondence with the data shown in this paper (one 600 bp *Bam*H I fragment is missing on the whole genome map). The success of this project has prompted the malaria genome consortium to recommend funding of whole genome mapping to assist in closure of chromosomes, as well as for verification of the final assembly.

In summary, we describe the construction of an ordered restriction map of *P. falciparum* chromosome 2 using optical mapping of genomic DNA. A combined approach using shotgun sequencing and optical mapping will enable sequence assembly and finishing of large and complex genomes.

ACKNOWLEDGMENTS

This work was supported by the Burroughs Wellcome Fund and the Naval Medical Research and Development Command work unit STEP C611102A0101BCX. The opinions and assertions herein are those of the authors and are not to be construed as official or as reflecting the views of the US Navy or naval service at large.

REFERENCES

Aston, C., C. Hiort, and D.C. Schwartz. 1998. Optical Mapping: An approach for fine mapping. *Methods in Enzymology*: in press.

Cai, W., H. Aburatani, D. Housman, Y. Wang, and D.C. Schwartz. 1995. Ordered restriction endonuclease maps of yeast artificial chromosomes created by optical mapping on surfaces. *Proc. Natl. Acad. Sci. USA* **92**: 5164-5168.

Cai, W., J. Jing, B. Irvin, L. Ohler, E. Rose, U. Kim, M. Simon, and D.C. Schwartz. 1998. High resolution restriction maps of bacterial artificial chromosomes constructed by optical mapping. *Proc. Natl. Acad. Sci. USA* **95**: 3390-3395.

Corcoran, L.M., K.P. Forsyth, A.E. Bianco, G.V. Brown, and D.J. Kemp. 1986. Chromosome size polymorphism in plasmodium falciparum can involve deletions and are frequent in nature parasite populations. *Cell* **44**: 87-95.

Foote, S.J. and D.J. Kemp. 1989. Chromosomes of malaria parasites. *Trends Genet.* **5**: 337-342.

Dame, J.B., D.E. Arnot, P.F. Bourke, D. Chakrabarti, Z. Christodoulou, R. L. Coppel, F. Cowman, A.G. Craig, K. Fischer, J. Foster, N. Goodman, K. Hinterberg, A. A. Holder, D. C. Holt, D. J. Kemp, M. Lanzer, A. Lim, C. I. Newbold, J. V. Ravetch, G. R. Reddy, J. Rubio, S. M. Schuster, X. Z. Su, J. K. Thompson, F. Vital, T.E Wellems, and E. B Werner. 1996. Current status of the *Plasmodium falciparum* genome project. *Molec. Biochem. Parasitol.* **79**:1-12.

Foster, J. and Thompson, J. 1995. The *Plasmodium falciparum* genome project: A resource for researchers. The Wellcome Trust Malaria Genome Collaboration. *Parasitol. Today* **11**:1-4.

Gardner et al. 1998. Manuscript in preparation.

Jing, J., J. Reed, J. Huang, X. Hu, V. Clarke, J. Edington, D. Housman, T. Anantharaman, E. Huff, B. Mishra, B. Porter, A. Shenker, E. Wolfson, C. Hiort, R. Kantor, C. Aston, and D.C. Schwartz. 1998. Automated High Resolution Optical Mapping Using Arrayed, Fluid Fixed, DNA Molecules. *Proc. Natl. Acad. Sci. USA* **95**: in press.

Lin, J., R. Qi, C. Aston, J. Jing, T.S. Anantharaman, B. Mishra, O. White, J.C. Venter, and D.C. Schwartz. 1998. Complete Shotgun Optical Mapping of *Deinococcus radiodurans* and *Escherichia coli* K12 using Genomic DNA Molecules. Submitted.

Netzel, T.L., K. Nafisi, M. Zhao, J.R. Lenhard, and I. Johnson. 1995. Base-content dependence of emission enhancements, quantum yields, and lifetimes for cyanine dyes bound to double-strand DNA: Photophysical properties of monomeric and bichromophoric DNA stains. *J. Phys. Chem.* **99**: 17936-17947.

Samad, A.H., W.W. Cai, X. Hu, B. Irvin, J. Jing, J. Reed, X. Meng, J. Huang, E. Huff, B. Porter, A. Shenker, T. Anantharaman, B. Mishra, V. Clarke, E. Dimalanta, J. Edington, C. Hiort, R. Rabbah, J. Skiadas, and D.C. Schwartz. 1995. Mapping the genome one molecule at a time - optical mapping. *Nature* **378**: 516-517.

Schwartz, D.C., and M. Koval. 1989. Conformational dynamics of individual DNA molecules during gel electrophoresis. *Nature* **338**: 520-522.

Schwartz, D.C., X. Li, L. Hernandez, S. Ramnarain, E. Huff, and Y. Wang. 1993. Ordered restriction maps of *Saccharomyces cerevisae* chromosomes constructed by optical mapping. *Science* **262**: 110-114.

Sinnis, P, and T.E. Wellems. 1988. Long-range restriction maps of *Plasmodium falciparum* chromosomes: Crossingover and size variation among geographically distant isolates. *Genomics* **3**:287-295.

Trager, W. and J.B. Jensen. 1976. Human malaria parasites in continuous culture. *Science* **193**: 673-675.

Thompson, J.K, and Cowman, A.F. A YAC contig and high resolution map of chromosome 3 from *Plasmodium falciparum*. 1997. *Molec. Biochem. Parasitol.* **90**:537-542.

Watanabe, J., and J. Inselburg. 1994. Establishing a physical map of chromosome No. 4 of *Plasmodium falciparum*. *Molec. Biochem. Parasitol.* **65**:189-99.

Optical Map (kb)	Map predicted from sequence (kb)	difference (kb)	relative difference	Hybridizing probe
71.8	66.597	5.24		
114.5	115.147	0.63	0.6%	
10.3	10.226	0.02	0.2%	
3.4	3.359	0.07	2.1%	
7.9	7.856	0.05	0.6%	
24.7	23.684	1.03	4.4%	
6.8	4.933	1.88	38.0%	
16.5	14.553	1.97	13.6%	
3.2	2.875	0.30	10.3%	
	0.177			
11.5	11.425	0.10	0.9%	
4.1	3.768	0.30	7.9%	
63.8	63.252	0.50	0.8%	
10.0	10.018	0.01	0.1%	
6.7	6.431	0.27	4.2%	
8.9	9.248	0.31	3.3%	
28.7	27.327	1.34	4.9%	
4.3	4.357	0.07	1.6%	
7.6	7.581	0.01	0.1%	PF2CM93
11.0	10.588	0.44	4.2%	-
60.5	60.324	0.21	0.4%	
12.3	11.935	0.40	3.3%	
4.1	3.964	0.12	3.0%	
58.2	57.925	0.25	0.4%	
5.5	5.381	0.07	1.3%	
	0.363			
1.6	1.546	0.02	1.5%	
23.4	22.405	0.96	4.3%	
35.1	34.171	0.91	2.6%	
18.1	17.156	0.93	5.4%	
3.1	2.947	0.16	5.4%	
24.9	25.138	0.28	1.1%	
40.8	40.107	0.73	1.8%	
20.8	20.176	0.59	2.9%	
25.1	24.476	0.62	2.5%	
77.3	75.172	2.15	2.9%	PF2NA66
16.6	16.637	0.07	0.4%	
48.0	45.683	2.30	5.0%	
9.4	8.546	0.88	10.3%	
20.1	18.986	1.15	6.0%	

23.9	23.192	0.75	3.2%	
32.1	14.897	5.65		
976.5	934.513	0.60	4.3%	

Table 1. Comparison of *Nhe* I optical map with restriction map predicted from sequence.

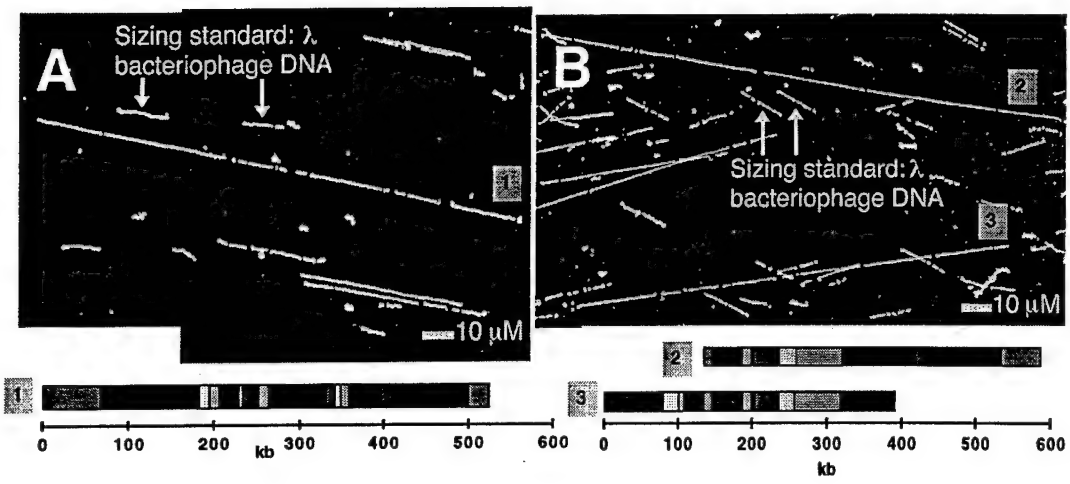
Optical Map (kb)	Map predicted from sequence (kb)	difference (kb)	relative difference	Hybridizing probe
77.1	76.648	0.42		
19.9	20.955	1.07	5.09%	
7.5	6.81	0.65	9.52%	
26.1	27.054	0.95	3.52%	
9.9	9.831	0.11	1.15%	
41.0	43.295	2.28	5.26%	
12.4	13.647	1.22	8.92%	
3.7	3.754	0.02	0.67%	
34.8	35.985	1.18	3.28%	
21.1	20.22	0.91	4.51%	
63.6	61.785	1.80	2.92%	
55.9	55.217	0.73	1.32%	
41.3	40.788	0.50	1.22%	PF2CM93
67.3	70.318	3.05	4.33%	
46.7	46.943	0.23	0.49%	
81.2	87.327	6.14	7.03%	
2.0	1.786	0.20	11.35%	
8.9	11.633	2.68	23.07%	
18.6	17.953	0.69	3.85%	
80.8	83.96	3.16	3.77%	
19.9	20.665	0.78	3.76%	
31.1	30.351	0.72	2.39%	
17.4	17.959	0.56	3.10%	
28.6	30.812	2.22	7.21%	PF2NA66
52.2	49.95	2.26	4.52%	
2.0	1.813	0.18	9.70%	
24.9	24.79	0.07	0.28%	
6.0	5.315	0.65	12.28%	
0.5	0.621	0.12	19.48%	
34.8	16.346	6.93		
937.2	934.531	1.25	5.86%	

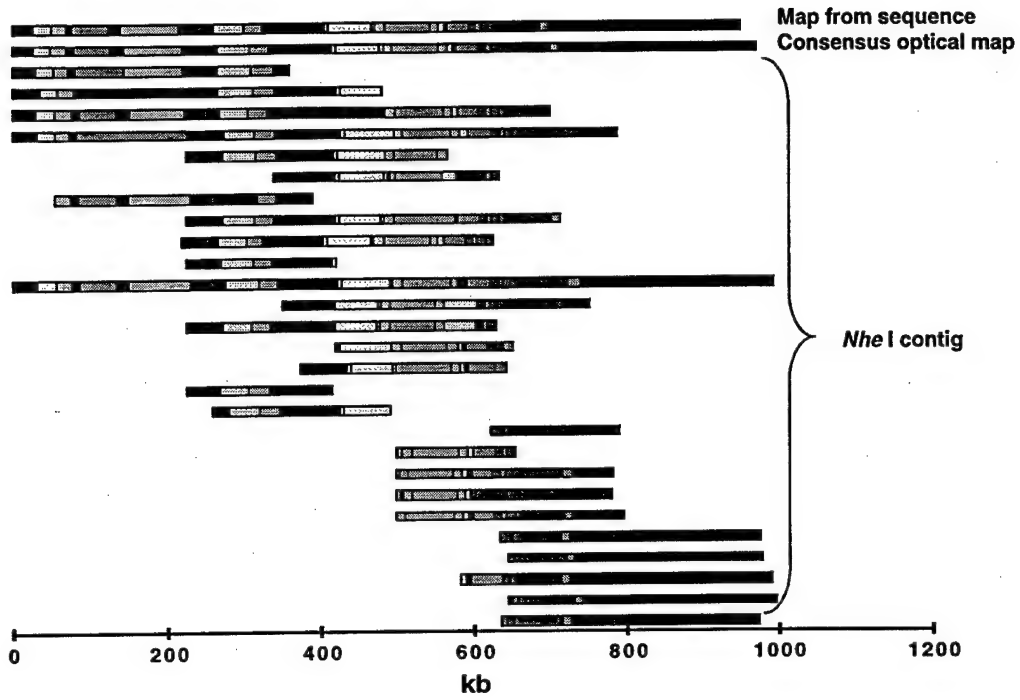
Table 2. Comparison of *Bam*H I optical map with restriction map predicted from sequence.

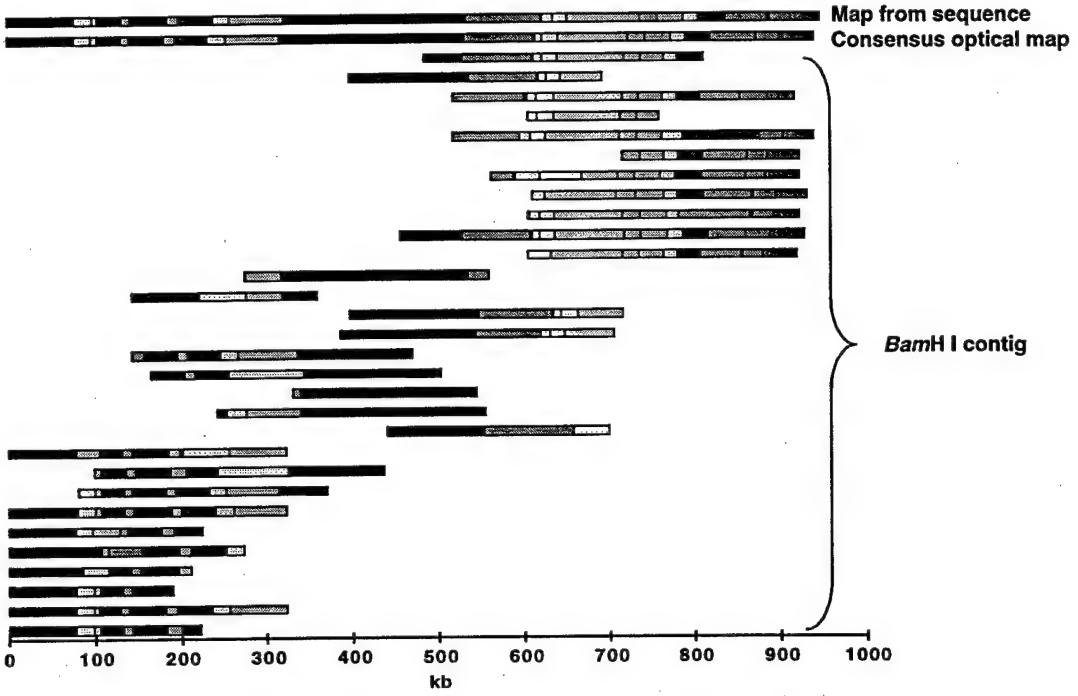
FIGURE LEGENDS

Fig. 1. Typical *P.falciparum* chromosome 2 molecules and their corresponding optical maps. (A) digested with *Nhe* I (B) digested with *Bam*H I. Maps derived from the two *Bam*H I-digested molecules in (B) can be aligned.

Fig. 2. High resolution Optical Mapping of *P. falciparum* chromosome 2 using (A) *Nhe* I and (B) *Bam*H I. The underlying contig used to generate the consensus map is shown. The map predicted from sequence information is shown for comparison.


A
Sizing standard: λ
bacteriophage DNA
1
10 μM
B
2
Sizing standard: λ
bacteriophage DNA
3
10 μM
1
0
100
200
kb
300
400
500
600
2
3
0
100
200
300
400
500
600
kb


Map from sequence
Consensus optical map
Nhe I contig
0
200
400
600
800
1000
1200
kb


Map from sequence
Consensus optical map
*Bam*H I contig
0
100
200
300
400
500
600
700
800
900
1000
kb

DRAFT

Interpolated Markov models for eukaryotic gene finding

Steven L. Salzberg
Corresponding author. The Institute for Genomic Research, 9712 Medical Center Drive, Rockville, MD 20850. Phone: 301-315-2537. Fax: 301-838-0209. Email: salzberg@tigr.org.

Mihaela Pertea
Department of Computer Science, Johns Hopkins University, Baltimore, MD 21210. Email: Mihaela.Pertea@loria.fr.

Arthur L. Delcher
Department of Computer Science, Loyola College in Maryland, Baltimore, MD 21210 and Celera Genomics, 45 W. Gude Dr., Rockville, MD 20850. Email: delcher@cs.loyola.edu.

Malcolm J. Gardner
The Institute for Genomic Research, 9712 Medical Center Drive, Rockville, MD 20850. Email: gardner@tigr.org.

Herve Tettelin
The Institute for Genomic Research, 9712 Medical Center Drive, Rockville, MD 20850. Email: tettelin@tigr.org.

Running title: IMMs for eukaryotic gene finding

Abstract

Computational gene finding research has emphasized the development of gene finders for bacterial and human DNA. This has left genome projects for some small eukaryotes without a system that addresses their need. This paper reports on a new system, GlimmerM, that was developed to find genes in the malaria parasite *P. falciparum*. Because the gene density in *P. falciparum* is relatively high, the system design was based on a successful bacterial gene finder, GLIMMER. The system was augmented with specially trained modules to find splice sites, and was trained on all available data from the *P. falciparum* genome. Although a precise evaluation of its accuracy is impossible at this time, laboratory tests (using RT-PCR) on a small selection of predicted genes confirmed 100% of those predictions. With the rapid progress in sequencing the genome of *P. falciparum* the availability of this new gene finder will greatly facilitate the annotation process.

1 Introduction

The gene finding research community has focused considerable attention on human gene finding and bacterial gene finding. This is not surprising given the attention paid to both areas. The Human Genome Project has produced many millions of nucleotides of sequence, and the importance of rapidly identifying the genes in this sequence cannot be overstated. This task is made difficult by the fact that only one to three percent of human genomic sequence is estimated to code for proteins. On the bacterial side, eighteen complete bacterial and archaeal genomes have already been published, with dozens more expected in the next two years. Gene finders for these prokaryotes have an advantage in that approximately 90% of the DNA of these genomes is coding; thus the task reduces in many cases to choosing between competing reading frames. On the other hand, the demand for accuracy is correspondingly much higher in the prokaryotic world.

In between these two genomic worlds lies a vast array of eukaryotic organisms whose genomes range in size from that of a large prokaryote (on the order of tens of millions of nucleotides) to those that are larger than human (billions of nucleotides). Their gene density tends to be much lower than that of bacteria, but many organisms have a much higher gene density than humans. The only eukaryotic genome that has been completely sequenced, that of the yeast *Saccharomyces cerevisiae*, has approximately one gene every five kilobases. This corresponds to a gene density of 20%. Recently, chromosome 2 of the malaria parasite *Plasmodium falciparum* was completed (Gardner et al. 1998), and this organism too has a gene density of 20%. The remaining 13 chromosomes from malaria should be completed over the course of the next few years. The much larger (120 million nucleotides) genome of arabidopsis, which also is expected to have a gene density of approximately 20%, should be completed in the same time frame, and many projects are under way to sequence other small eukaryotes.

Because of their relatively high gene density with respect to human DNA, using a gene finder developed for human sequence (or other organisms with low gene density, including most vertebrates and larger plant genomes) may not be the optimal approach for *P. falciparum* and other small eukaryotes. Prokaryotic gene finders are not well-suited to this task because of their inability to handle introns. It is possible to re-train human gene finders using different data (for example, GENSCAN (Burge and Karlin 1997) has been trained with arabidopsis data), but one still runs the risk that because these systems have been optimized to find genes in DNA that is only 3% coding, they may miss many genes in genomes such as *P. falciparum*.

This paper describes a gene finder developed specifically for small eukaryotes with a gene density of around 20%. This system, GLIMMERM, was built and trained using data from *P. falciparum*, the malaria parasite. It was then used as the principal gene finder for chromosome 2 of *P. falciparum*, which contains 210 genes (209 protein coding genes plus one tRNA) (Gardner et al. 1998). Most of these genes were found by GLIMMERM, and as described below, some were confirmed by additional laboratory experiments.

2 Methods and algorithms

The basis of GLIMMERM is a dynamic programming algorithm that considers all combinations of possible exons for inclusion in a gene model, and chooses the best of these combinations. The decision about what model is best is a combination of the strength of the splice sites and the score of the exons produced by an interpolated Markov model (IMM). The methods for producing these scores are described next, followed by the description of the dynamic programming algorithm and the splice site recognition procedures.

2.1 Interpolated Markov models

Markov chains are a family of methods for computing the probability of an event based on a fixed number of previous events. In the context of DNA sequence analysis, Markov chains predict a base by examining a fixed number of bases just prior to that base in the sequence. The most common type of Markov chain is a fixed-order chain, in which the number of previous bases to examine is specified in advance. For example, a 5th-order Markov chain will predict a base by looking at the five previous bases. Markov chains, and 5th-order chains in particular, have proven to be effective at gene prediction in bacterial genomes (Borodovsky and Mcininch 1993; Borodovsky et al. 1995).

Interpolated Markov models (IMM) are an improvement on fixed-order Markov chains. The main distinction is that rather than deciding in advance how many bases to consider for each prediction, these models will use varying numbers of bases for each prediction. In some contexts they will use 5 bases, while in others they might use 6 or more bases, and in yet other cases they may use 4 or fewer bases. This allows IMMs to be sensitive to how common a particular oligomer is in a given genome. In a given genome, many 5-mers might occur rarely and should not be used for prediction; here the IMM will fall back on a shorter Markov chain. On the other hand, certain 8-mers may occur very frequently, and for those the IMM can use this longer context and make a better prediction. In addition, the IMM can combine the evidence from the 8^{th}-order Markov chain and the 5^{th}-order chain in such cases. Thus it has all the information available to a 5^{th}-order chain plus additional information. It is also worth noting both both IMMs and 5^{th}-order Markov chains should outperform methods based on codon usage statistics. (Cf. (Saul and Battistutta 1988), a codon usage method specific to *P. falciparum*. Note that at the time of that work, much less Plasmodium data was available, and higher-order statistics might have been inaccurate as a result.)

IMMs form the basis of the GLIMMER system for finding genes in bacteria and archaea

(Salzberg et al. 1998). GLIMMER correctly identifies approximately 98% of the genes in bacteria without any human intervention, and with a very limited number of false positives. It has been used as the gene finder for *B. burgdorferi* (Fraser et al. 1997), *T. pallidum* (Fraser et al. 1998), *C. trachomatis* (Stephens et al. 1998), *T. maritima* (*et al.* 1999,), and others. Based on the success of GLIMMER in bacterial sequence annotation, we thought that IMMs should make a good foundation for eukaryotic gene finding. This is particularly true of small eukaryotes like *P. falciparum* in which the gene density is intermediate between that of prokaryotes and higher eukaryotes.

Details of how to construct an IMM for sequence data can be found in the original GLIMMER publication (Salzberg et al. 1998); GLIMMERM uses the same IMM algorithm as the one described there. In brief, GLIMMERM builds IMMs from a set of DNA sequences chosen for training. For coding regions, it builds 3 separate IMMs, one for each codon position. (This is known as a 3-periodic Markov model (Borodovsky and Mcininch 1993).) These IMMs include 0^{th} through 8^{th} order Markov chains, as well as weights computed for every oligomer of 8 bases or less that appears in the training data. These weights and Markov models are interpolated to produce a score for each base in any potential coding sequence. The logs of these scores are summed to score each coding region.

2.2 Dynamic programming

GLIMMERM uses dynamic programming (DP) to find genes in the DNA sequence of a malaria chromosome. DP allows it to prune out a large number of possible exon-intron combinations and focus its analysis only on relatively high-scoring combinations (called "parses"). The input to the algorithm is any genomic DNA sequence in a FASTA format; small sequences as well as entire chromosomes can be input. The output is a partitioning of the the DNA into coding regions interleaved with noncoding regions, on

both the main and complementary strands of the sequence.

Wrapped inside GLIMMERM is the interpolated Markov model described above, which is used to score candidate exons. The predicted genes are optimal with respect to the scores produced by the IMM.

Dynamic programming has been the basis of many successful eukaryotic gene finders. Hidden Markov model (HMM) systems use a DP algorithm called Viterbi that is a special case of the algorithm here; these HMM methods include VEIL (Henderson et al. 1997); GENSCAN (Burge and Karlin 1997), which uses semi-Markov HMMs;and Genie (Kulp et al. 1996), which uses generalized HMMs. Very recently, Wirth descrbed a gene finder for *P. falciparum* based on generalized HMMs (Wirth 1998), but it is not yet available for comparison. The Morgan system (Salzberg et al. 1996; Salzberg et al. 1998) uses a DP algorithm as a wrapper around its decision tree program, and GeneParser (Snyder and Stormo 1995) uses DP wrapped around a neural network program. These latter two DP formulations are most similar to the one used for GLIMMERM.

As in many other gene finders (Salzberg 1998), there are a number of assumptions used by GLIMMERM when predicting genes in the DNA sequence. These assumptions are derived from biological constraints on mRNA splicing, transcription, and translation. Consequently, we assume that:

- each coding region of a gene begins with a start codon ATG,
- a gene has exactly one in-frame stop codon, which appears as the last codon in the gene
- each exon must be in the same reading frame as the previous exon
- every intron begins with the dinucleotide GT and ends with the dinucleotide AG.

These constraints significantly enhance the efficiency of computing the optimal parse of the DNA sequence, by restricting the search space of the DP algorithm. On the other

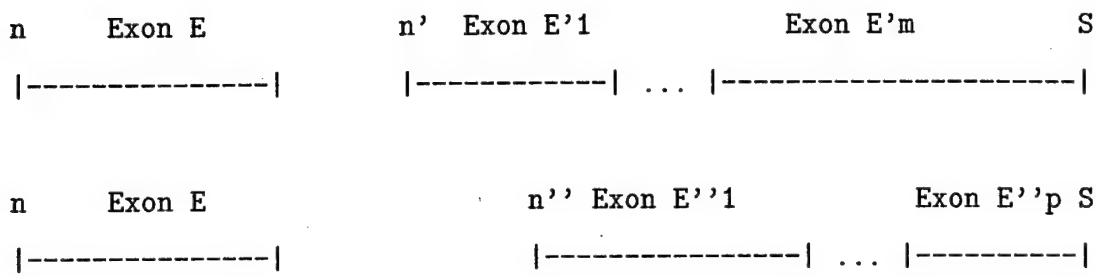


Figure 1: Dynamic programming to build gene models.

hand, genuine frameshifts cannot be detected by the system.

The dynamic programming algorithm starts building a putative gene after finding a stop codon in one of the six possible reading frames (considering both the forward and reverse directions). It attempts to complete the gene by searching backwards from the stop codon to the various possible start codons, adding exons to the model along the way.

After identifying all potential splice sites (described below), the DP algorithm then finds potential exons. At each potential start codon or acceptor site n, the algorithm decides if there is an exon starting at n. The algorithm prefers to make the exon as long as possible; this is generally a very useful heuristic for *P. falciparum*, in which the 82% AT-content makes long stretches of noncoding DNA without a stop codon very unlikely. After delimiting the exon, the algorithm searches downstream for all optimal parses ending at the same stop codon and consistent with the same reading frame of the exon under consideration.

The algorithm is best illustrated by example, as shown in Figure 1. Let S denote the stop codon, and $P(n, S)$ the optimal parse of the sequence starting at location n, and ending at the stop codon S. Suppose we are at location n and we have identified exon E. We want to find $P(n, S)$. Because the algorithm proceeds backwards from the stop codon, it has already computed $P(n', S)$, an optimal parse starting at location n' and

ending at S, which is made of the exons $E'1, ..., E'm$. It has also identified $P(n'', S)$ an optimal parse starting at location n'' and ending at S, made up of the exons $E''1, ..., E''p$. In the case of a gene on the direct strand, $n < n'$ and $n < n''$. Both $P(n', S)$ and $P(n'', S)$ are in the same reading frame with exon E. To find an optimal parse $P(n, S)$, we use the IMM obtained in the training phase to score the sequence obtained by concatenating E with $E'1, ..., E'm$. Likewise we score the sequence obtained after concatenating E and $E''1, ..., E''p$. GLIMMERM will choose the region that scores best as an optimal parse $P(n, S)$. In the case of equal scores, the system prefers the longest gene model.

A final rule inserted into the GLIMMERM calculations was the use of AT-content to identify exons. It has been observed that *P. falciparum* exons have an AT-content that averages around 70-75%, compared to the genome average of 82%. Thus when deciding among alternative gene models, this statistic sometimes provides additional help. This rule was incorporated into a post-processing step.

After predicting gene structures in both direct and complementary strands, GLIMMERM makes one more pass over the sequence to reject overlapping genes. Two overlapping genes are scored again using the IMM, and the one with the better score is retained. Gene models have a minimum total length of 200bp (computed as the the sum of all coding bases); this value can be easily adjusted by the user.

In order to avoid mistakenly "committing" to the wrong gene model too soon in its calculations, GLIMMERM was designed to output several competing models, along with their scores, for a given subsequence. This allowed the human annotators to choose among models using other information (such as alignment data) when such information was available.

2.3 Splice site identification

The approach used by GLIMMERM to determine the splice sites is similar to the one described in (Salzberg et al. 1998). A 2nd-order Markov chain model is used to score a 16-base region around donor sites and a 29-base region around acceptor sites. Two 2nd-order Markov models were built for each type of site. First, a "true" Markov model was created from existing data on known 5' and 3' consensus sites. This data was collected by exhaustively combing the literature for every documented exon-intron boundary. A "false" Markov model was built from a large number of randomly chosen false splice sites; i.e., sequences that contained the consensus GT or AG dinucleotide but that were not true splice sites. The score of a site $s_i, s_{i+1}, \ldots, s_j$ was computed by each Markov model according to the formula:

$$S(i,j) = \sum_{k=i}^{j} M_{s,k}$$

where

$$M_{s,k} = \ln\left(f_{(s_{k-2},s_{k-1},s_k),k} / f_{(s_{k-2},s_{k-1}),k-1}\right)$$

and $f_{s,k}$ is the frequency of substring s ending at location k. Note that for the leftmost position in the splice site region, M is taken to be the probability given by the 0^{th} order Markov model, and for the second position, M is given by the 1^{st} order model. The score for a given splice site is computed by taking the difference of the scores obtained from the "true" site Markov model and the "false" site model.

After building the models, we then scored all the true splice sites and a large selection of randomly chosen false sites. We then set minimum cutoff scores in order to correctly identify most (or all) true sites, and measured how many false positives we would expect with various thresholds.

Figure 2 shows the tradeoff in thresholds for the splice site recognition function in *P. falciparum*. The figure reflects the state of the system after re-training in late 1998

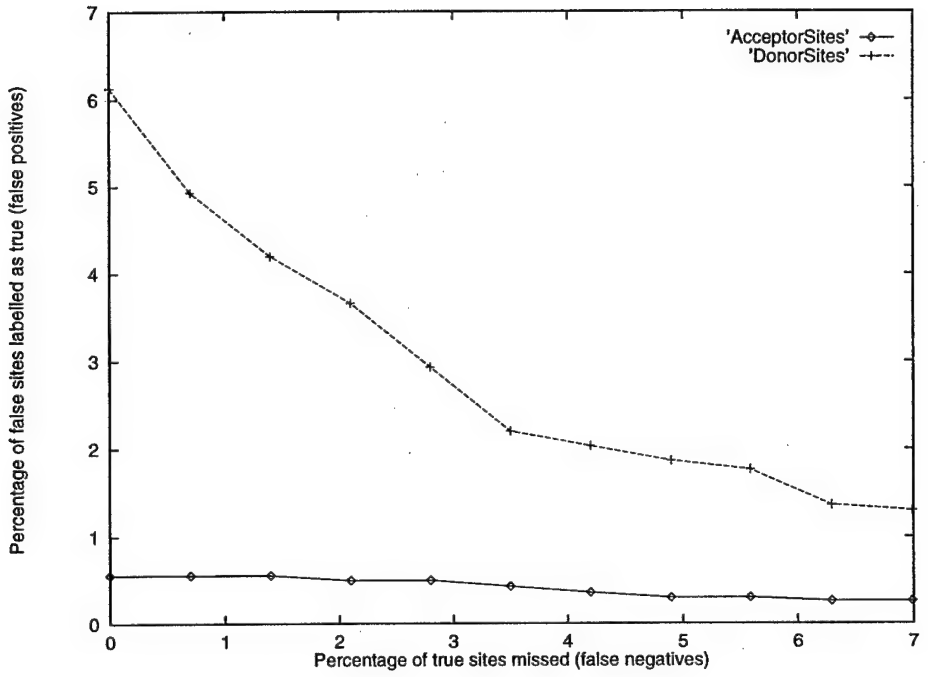


Figure 2: Tradeoff between false positive rates and false negative rates for the Markov chain method that recognizes exon-intron splice sites.

(after the conclusion of the chromosome 2 sequencing project), at which time 143 introns were available from *P. falciparum*. The figure shows the tradeoff between sensitivity and selectivity for the Markov chain method on both donor and acceptor sites. Acceptor sites are much easier to recognize: with a false negative rate of 0% (corresponding to a sensitivity of 100%, meaning that all true sites will be recognized), the false positive rate (the percentage of AG dinucleotides that will incorrectly be called acceptor sites) is just 0.56%. For donor sites, a 0% false negative rate corresponds to a rather high 6.1% false positive rate. Setting the system so that it misses 4 of the 143 true donor sites (2.8%) reduces this false positive rate to 2.9%.

Note that before sequencing of chromosome 2, only about 90 introns were known and the Markov chain models were consequently not as accurate as those here. The implication is that GLIMMERM will perform even better for subsequent chromosomes of malaria, and we intend to continue re-training it as the genome sequencing progresses. The additional introns were identified with high confidence by aligning the genomic sequence of chromosome 2 with EST sequences that spanned introns and with protein sequences from other organisms. The statistics presented in the results section represent the performance of GLIMMERM after re-training to improve the donor and acceptor site Markov models.

2.4 Code availability

The complete source code for GLIMMERM will be made available soon; it has already been shared with other malaria genome sequencing centers. The code includes routines for re-training the system on data from other organisms. A version of the system trained on *Arabidopsis thaliana* genes is currently under development. Total processing time to find all genes in malaria chromosome 2 (approximately one million nucleotides) is about half an hour on a Pentium 450 personal computer running Linux.

2.5 Annotating a genome

In its current form, GLIMMERM produces multiple gene models for some genes. A model is a set of exons that, when concatenated together, form a single protein. When no database matches were found to support a GLIMMERM prediction, the chromosome 2 annotation reflects the highest scoring model. Although many of these are likely to be correct, it is undoubtedly the case that some are not. Further investigation is required to confirm these predictions (but see below for laboratory evidence confirming a small subset).

The GLIMMERM algorithm was used as one of a suite of tools. Accurate gene identification depends on using every tool available, and the description here should not be taken as implying that GLIMMERM alone can find all genes in *P. falciparum* or any other genome. However, it was a central component in a larger strategy. Other important computational tools used by the malaria chromosome 2 team were: (1) searches of a nonredundant protein sequence database using gapped BLAST and PSI-BLAST (Altschul et al. 1990; Altschul et al. 1997); (2) gapped alignments of DNA to protein and EST sequence databases using DDS and DPS (Huang et al. 1997); (3) prediction of putative signal peptides using SignalP (Nielsen et al. 1997); (4) prediction of transmembrane domains with PHThtm (Rost et al. 1995); and (5) prediction of nonglobular structures with SEG (Wootton and Federhen 1996). In addition, the project used additional aligment tools developed at TIGR to detect frameshift errors: these tools allow an annotator to detect when a sequence alignment extends beyond the start and stop codons indicated by other tools. In some cases this indicates errors in sequencing, which can be corrected; in other cases it indicates either a genuine frameshift that occurs during translation or a mutation that has changed the length of the translated protein. Any comprehensive annotation effort needs these computational tools and more in order to produce reasonably accurate gene annotations.

3 Results and discussion

GLIMMERM was used as the primary gene finder for chromosome 2 of *P. falciparum*, for which it finds 207/209 (99%) genes automatically, as detailed below. In addition to GLIMMERM, the annotation process used other computational tools as mentioned above. Chromosome 2 has 209 protein-coding genes spread over approximately one million bases, for a gene density of one gene per 4.5 kb. This contrasts with a density of 1/kb in bacteria, 1/2kb in yeast, 1/7kb in *C. elegans*, and 1/50kb (estimated) in human. Of the 209 protein-coding genes, 43% had at least one intron and those genes with introns usually had just one or two introns (Gardner et al. 1998).

3.1 Training

In order to train the IMM, we needed to collect as much coding sequence as possible from *P. falciparum* itself. We exhaustively surveyed the literature to collect every sequence, including partial genes, that was backed by laboratory evidence. Our survey collected 110 complete coding sequences from all 14 chromosomes, of which just 6 came from chromosome 2. (Note that by length, chromosome 2 comprises approximately 3% of the genome.) This training set provided the data for the splice site models described above as well.

An important point to emphasize here is that *P. falciparum* has an unusually high 82% AT content. As a consequence of this high AT content, stop codons are very frequent (e.g., TAA will occur especially often) in noncoding DNA. This makes it much more likely that long open reading frames (ORFs) represent coding sequence. This fact was used to generate additional training data for GLIMMERM: ORFs greater than 500bp in the chromosome 2 sequence were assumed to be coding regions, and were used in the IMM training. These were added to the list generated by the literature search.

3.2 Accuracy

Of the 209 genes, GLIMMERM finds 178 exactly; i.e., it identifies the correct start codon, the correct boundaries of every exon and intron, and the correct stop. Of these, 40 have competing gene models that score higher, meaning that a human annotator had to examine the output and decide which gene model looked best. This process was frequently made easier by the existence of EST or protein matches.

Of the remaining 31 genes, GLIMMERM finds the stop codons correctly for 14 of these. Different starts appear in the final annotation for several reasons; for example, the existence of a match to a protein sequence that starts at a different start codon. (Note that it is entirely possible that GLIMMERM is still correct in these cases.) In rare cases, a protein hit produces a coding region that contains a stop codon; these indicate genuine frameshifts and are annotated as such in the database. The system finds the correct start but the wrong stop codon for four genes; this occurs in multi-exon genes in which a splice site is missed and one of the exons is incorrectly extended until it hit a stop codon. The 11 remaining partial hits are cases in which GLIMMERM predicts some but not all exons correctly; for example, three multi-exon genes are each broken into two separate genes.

Only two of the 209 genes are missed completely. One is a predicted integral membrane protein of 192aa predicted by the original version of GLIMMERM (before retraining the splice site models). A separate program was used to predict the function of this protein; it did not align to any known sequences. The second is ribosomal protein S30; ribosomal proteins often have a strikingly different composition from other genes and are known to be difficult for content-based gene finders to locate. These will not be missed as long as genomic data is searched against databases of known ribosomal proteins.

The improved splice site Markov models resulted in GLIMMERM's generating 41

fewer gene models than before. In addition to the one missed gene just described, it generated five new gene models. Of these, one appears to be a secreted protein, and we are currently investigating this to see if it should be added to the published annotation.

A significant caveat to include with these results is that GLIMMERM often produces multiple competing models that the human annotator must resolve. Most genes with three or more exons result in multiple models. The system indicates which model scores the highest, but as indicated above, 40 of the "correct" gene models had alternative parses that scored higher. These alternative parses share some exons but use different splice sites for others. A human annotator, looking at additional evidence, was able to overrule the system's top choice in these cases. It is likely that in other cases where no evidence besides GLIMMERM's prediction is available, some of the published annotation may still be in error (all such proteins are annotated as hypotheticals). After collapsing each set of multiple gene models into one model, the gene list still contains 266 genes. These means that (since only 209 genes appeared in the final annotation) the annotators eliminated another 57 gene models entirely from the output. These decisions were somewhat subjective: frequently the putative genes were short, or they consisted mostly of low-complexity sequence, and this was not enough to convince the human annotators that the genes were real. In many cases the annotators are probably correct, but it is simply impossible at this point to say with confidence that all of the deleted genes are false positives. Only further evidence will allow us to decide, but this makes clear the importance of continuing to update and improve genome annotation over time.

3.3 Performance on known genes

Another way to assess the accuracy of the program is to consider its accuracy on those proteins whose exon-intron structure is known precisely from laboratory studies. There are seven genes from chromosome 2 of *P. falciparum* that currently fit into this category;

Table 1: Performance of GLIMMERM on genes whose structure is completely known from independent laboratory evidence. All seven genes had perfect matches to the system's predictions, meaning that the start codon, stop codon, and every splice site were correctly predicted. The column headings give the gene name, its length in amino acids, number of introns (Intr), a comment on GLIMMERM's prediction, and the common name of the protein.

Name	Len	Intr	Comment	Common name
PFB0100c	654	1	Perfect match	knob-associated His-rich prt
PFB0295w	471	0	Perfect match	adenylosuccinate lyase (OO)
PFB0300c	272	0	Perfect match	merozoite surface antigen MSP-2
PFB0305c	272	1	Perfect match	merozoite surface antigen MSP-5 (EGF domain)
PFB0310c	272	1	Perfect match, highest score from 5 models	merozoite surface antigen MSP-4 (EGF domain)
PFB0340c	997	3	Perfect match, differed at one splice site (see main text)	SERA antigen/papain-like protease with active Ser
PFB0405w	3135	0	Perfect match, higher score from 2 models	transmission blocking target antigen PfS230

i.e., the sequence from start to stop has been completely characterized. Of these seven, one (PFB0100c) was published subsequent to the construction of our training set; the other six were included in the training data.

GLIMMERM's performance on this small set of genes is shown in Table 1. For all seven of the genes, GLIMMERM's output contained a model that matched perfectly. For four of the genes, there correct model was the only one output by the system. For PFB0310c and PFB0405c, GLIMMERM produced five and two competing models respectively, but in each case the highest scoring one was correct. Only for PFB0340c, a 4-exon gene, was GLIMMERM's highest scoring model not the right one. The system did produce the correct answer, but it gave a slightly higher score to a model that used a different donor site for the first exon. GLIMMERM's alternate prediction would have a 23aa insertion in this 997aa protein.

3.4 Laboratory tests

The only way of measuring the accuracy of GLIMMERM precisely is to test each of its predictions in the laboratory to see if they are expressed as predicted. Alternatively, for those genes with significant homology to genes from other organisms, the homologous gene can be used as an independent confirmation of the prediction. For chromosome 2, 90 predicted proteins (43%) have no homolog in the existing databases (Gardner et al. 1998), and the task of testing all of these predictions has not yet been conducted. Over time, we expect some of them to show up as homologs, obviating the need for laboratory experiments. However, one careful set of experiments was conducted as part of the chromosome 2 study.

Because many of the proteins predicted by GLIMMERM had unusual nonglobular domains, the chromosome 2 project team ran a reverse transcriptase (RT-PCR) experiment for 13 of these genes (Gardner et al. 1998) to determine whether or not they were real. The RT-PCR focused its attention on nonglobular domains, not entire proteins, so it could not confirm every detail of the GLIMMERM predictions. This experiment confirmed that all 13 of the nonglobular domains were expressed; i.e., the predictions for those regions were correct. To our knowledge, this is the first time ever that computational predictions provided the impetus for experiments which in turn confirmed the predictions.

Acknowledgements

SLS is supported by the National Human Genome Research Institute at NIH under Grant No. K01-HG00022-1. SLS, ALD, and MP are supported in part by the National Science foundation under Grant No. IRI-9530462. MJG and HT were supported by a supplement to NIAID grant R01 AI40125-01, which was made possible with funds from NIH's Office for Research on Minority Health and Department of the Army Cooperative

Agreement DAMD17-98-2-8005.

References

Altschul, S., W. Gish, W. Miller, E. Myers, and D. Lipman (1990). Basic local alignment search tool. *J. Mol. Biol. 215*, 403–410.

Altschul, S., T. Madden, A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. Lipman (1997). Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic Acids Res. 25*(17), 3389–3402.

Borodovsky, M. and J. Mcininch (1993). Genemark: Parallel gene recognition for both DNA strands. *Computers & Chemistry 17*(2), 123–133.

Borodovsky, M., J. McIninch, E. Koonin, K. Rudd, C. Medigue, and A. Danchin (1995). Detection of new genes in the bacterial genome using Markov models for three gene classes. *Nucleic Acids Research 23*, 3554–3562.

Burge, C. and S. Karlin (1997). Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol. 268*, 78–94.

Fraser, C., S. Casjens, W. Huang, G. Sutton, R. Clayton, R. Lathigra, O. White, K. Ketchum, R. Dodson, E. Hickey, M. Gwinn, B. Dougherty, J.-F. Tomb, R. Fleischmann, D. Richardson, J. Peterson, A. Kerlavage, J. Quackenbush, S. Salzberg, M. Hanson, R. van Vugt, N. Palmer, M. Adams, J. Gocayne, J. Weidman, T. Utterback, L. Watthey, L. McDonald, P. Artiach, C. Bowman, S. Garland, C. Fujii, M. Cotton, K. Horst, K. Roberts, B. Hatch, H. Smith, and J. Venter (1997, December). Genomic sequence of a lyme disease spirochaete, *Borrelia burgdorferi. Nature 390*(6660), 580–586.

Fraser, C., S. Norris, G. Weinstock, O. White, G. Sutton, R. Clayton, R. Dodson, M. Gwinn, E. Hickey, K. Ketchum, E. Sodergren, J. Hardham, M. McLeod,

S. Salzberg, H. Khalak, J. Weidman, J. Howell, M. Chidambaram, T. Utterback, L. Watthey, L. McDonald, P. Artiach, C. Bowman, S. Garland, C. Fujii, M. Cotton, K. Horst, K. Roberts, B. Hatch, H. Smith, and J. Venter (1998). Complete genomic sequence of *Treponema pallidum*, the syphilis spirochete. *Science 281*, 375–388.

Gardner, M., H. Tettelin, D. Carucci, L. Cummings, L. Aravind, E. Koonin, S. Shallom, T. Mason, K. Yu, C. Fujii, J. Pederson, K. Shen, J. Jing, C. Aston, Z. Lai, D. Schwartz, M. Pertea, S. Salzberg, L. Zhou, G. Sutton, R. Clayton, O. White, H. Smith, C. Fraser, M. Adams, J. Venter, and S. Hoffman (1998). Chromosome 2 sequence of the human malaria parasite *Plasmodium falciparum. Science 282*, 1126–1132.

Henderson, J., S. Salzberg, and K. Fasman (1997). Finding genes in human DNA with a hidden Markov model. *Journal of Computational Biology 4*(2), 127–141.

Huang, X., M. Adams, H. Zhou, , and A. Kerlavage (1997). A tool for analyzing and annotating genomic sequences. *Genomics 46*, 37–45.

Kulp, D., D. Haussler, M. G. Reese, and F. H. Eeckman (1996). A generalized hidden Markov model for the recognition of human genes in DNA. In *ISMB-96: Proc. Fourth Intl. Conf. Intelligent Systems for Molecular Biology*, Menlo Park, CA, pp. 134–141. AAAI Press.

Nielsen, H., J. Engelbrecht, S. Brunak, and G. von Heijne (1997). Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Engineering 10*(1), 1–6.

Rost, B., R. Casadio, P. Fariselli, and C. Sander (1995). Transmembrane helices predicted at 95% accuracy. *Protein Science 4*(3), 521–533.

Salzberg, S. (1998). Decision trees and markov chains for gene finding. In S. Salzberg,

D. Searls, and S. Kasif (Eds.), *Computational Methods in Molecular Biology*, New Comprehensive Biochemistry, pp. 187–203. Amsterdam: Elsevier Science B.V.

Salzberg, S., X. Chen, J. Henderson, and K. Fasman (1996). Finding genes in DNA using decision trees and dynamic programming. In *ISMB-96: Proc. Fourth Intl. Conf. Intelligent Systems for Molec. Bio.*, Menlo Park, CA, pp. 201–210. AAAI Press.

Salzberg, S., A. Delcher, K. Fasman, and J. Henderson (1998). A decision tree system for finding genes in DNA. *J. Computational Biol. 5*(4), 667–680.

Salzberg, S., A. Delcher, S. Kasif, and O. White (1998). Microbial gene identification using interpolated Markov models. *Nucl. Acids Res. 26*(2), 544–548.

Saul, A. and D. Battistutta (1988). Codon usage in Plasmodium falciparum. *Molecular and Biochemical Parasitology 27*, 35–42.

Snyder, E. E. and G. D. Stormo (1995). Identification of coding regions in genomic DNA. *Journal of Molecular Biology 248*, 1–18.

Stephens, R., S. Kalman, C. Lammel, J. Fan, R. Marathe, L. Aravind, W. Mitchell, L. Olinger, R. Tatusov, Q. Zhao, E. Koonin, and R. Davis (1998). Genome sequence of an obligate intracellular pathogen of humans: Chlamydia trachomatis. *Science 282*(5389), 754–759.

et al., K. N. (1999). The complete genome sequence of *Thermitoga maritima*: Evidence for lateral transfer. *under submission*.

Wirth, A. (1998). A Plasmodium falciparum genefinder.

Wootton, J. and S. Federhen (1996). Analysis of compositionally biased regions in sequence databases. *Methods in Enzymology 266*, 554–71.